Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

Rieloff, Ellen

2021

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Rieloff, E. (2021). *Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins*. Lund University (Media-Tryck).

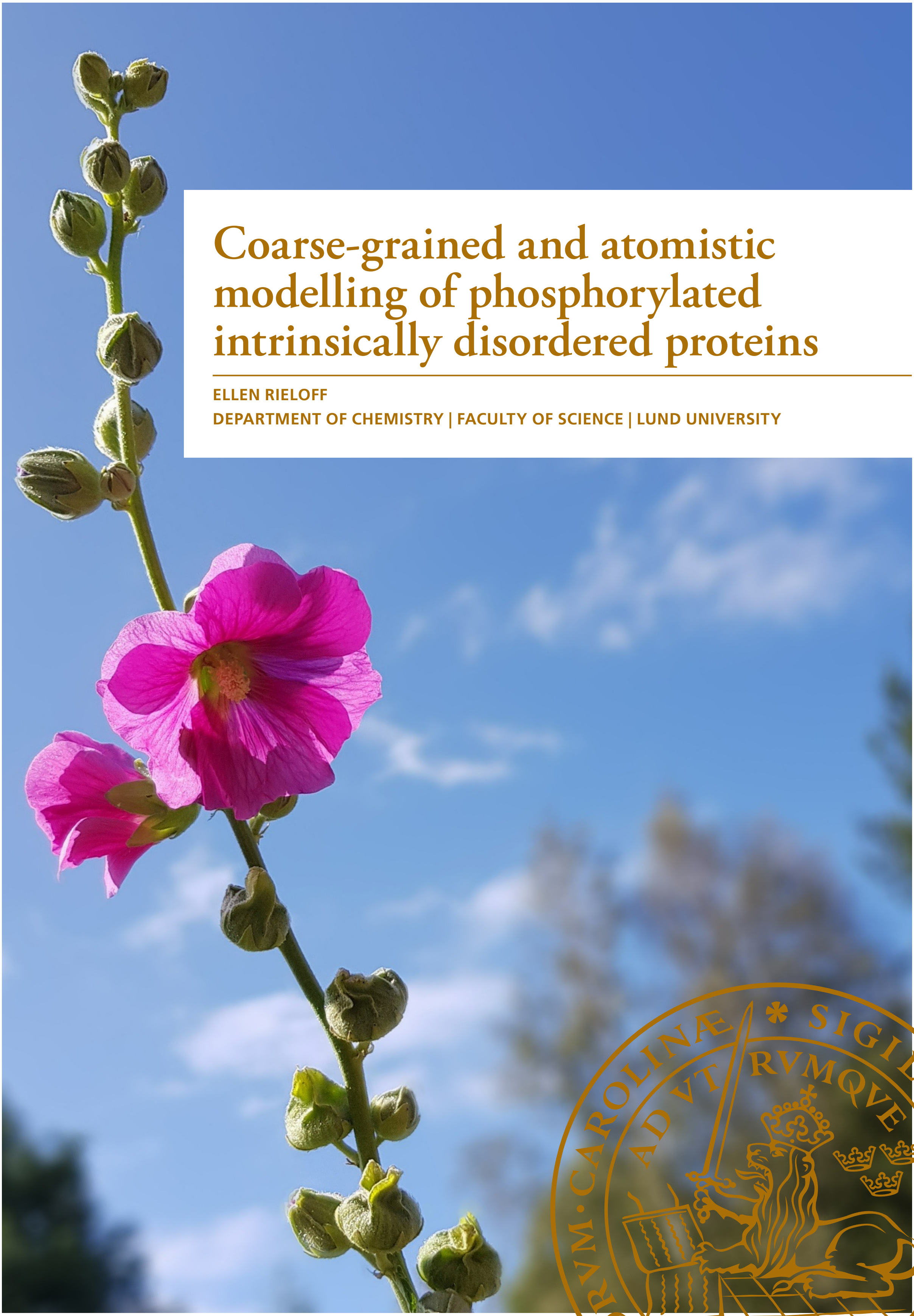
Total number of authors:
1

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

ELLEN RIELOFF
DEPARTMENT OF CHEMISTRY | FACULTY OF SCIENCE | LUND UNIVERSITY



Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

by Ellen Rieloff

DOCTORAL THESIS

by due permission of the Faculty of Science of Lund University, Sweden. To be defended on Friday, the 29th of October 2021 at 13:00 in lecture hall A at Kemicentrum.

Faculty opponent
Assoc. Prof. Elena Papaleo
Technical University of Denmark, Lyngby, Denmark.

Organization
LUND UNIVERSITY
Department of Chemistry

Document name
DOCTORAL DISSERTATION

Date of disputation
2021-10-29

Sponsoring organization

Author(s)
Ellen Rieloff

Title and subtitle
Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

Abstract

Intrinsically disordered proteins (IDPs) are involved in many biological processes such as signalling, regulation and recognition. One of the main questions regarding IDPs is how sequence, structure and function are related. Phosphorylation, a type of post-translational modification prevalent in intrinsically disordered proteins and regions, is an example of how modifications at the sequence level can induce changes in structure and thereby influence function. The lack of well-defined tertiary structure in IDPs makes them better described by an ensemble of conformations than a single structure. Furthermore, it causes them to be more difficult to study than conventional proteins, so a combined approach of experimental and simulation techniques are often advantageous. However, simulations rely on appropriate models. In this thesis, the conformational ensembles of IDPs, especially the saliva protein statherin, have been investigated using both simulations with different models and the experimental techniques small-angle X-ray scattering and circular dichroism spectroscopy. The aims have been to contribute to the collection of available tools for studying IDPs, by investigating models, and to explore the link between sequence and structure of IDPs, with special focus on phosphorylation. It was shown that a coarse-grained ”one bead per residue model” can be used to describe several different IDPs and provide an understanding of how protein length, charge distribution and salt concentration affects IDPs. Furthermore, by including a hydrophobic interaction the model could qualitatively describe the self-association of statherin and provide insight on the balance of interactions and entropy governing the process. The model was however shown to overestimate the compactness of longer and more phosphorylated IDPs. Turning to atomistic simulations, it was revealed that the conformational ensembles of phosphorylated IDPs are highly influenced by salt bridges forming between phosphorylated residues and arginine/lysine/C-terminus, such that over-stabilised salt bridges cause larger compaction than observed in experiments. Another force field could however detect phosphorylation-induced changes in global compaction and secondary structure and relate them to interactions between specific residues, illustrating the potential ability of simulations to provide insight into phosphorylation.

Key words
intrinsically disordered proteins, phosphorylation, simulations, Monte Carlo, molecular dynamics, coarse-graining, atomistic, statherin, small-angle X-ray scattering, circular dichroism

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-7422-828-1 (print)
978-91-7422-829-8 (pdf)

Recipient’s notes

Number of pages 274

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Ellen Rieloff

Date 2021-09-20

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

by Ellen Rieloff

LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarises the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Front cover: Photo by Ellen Rieloff.

Parts of this thesis has been published before in:
Rieloff, Ellen, *Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS* (2019)

Faculty of Science, Department of Chemistry

ISBN: 978-91-7422-828-1 (print)
ISBN: 978-91-7422-829-8 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

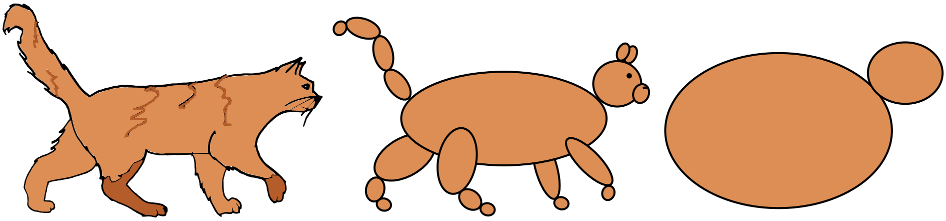
Till Ludvig
(Hoppas du gillar katten)

Contents

Populärvetenskaplig sammanfattning på svenska

Proteiner är en livsnödvändig komponent i våra kroppar. Dels är de viktiga byggstenar eftersom de ingår i kroppens alla vävnader, muskler och benstomme, men de har också andra kritiska uppgifter, såsom att transportera näringsämnen och syre samt försvara oss mot virus och bakterier. Länge trodde man att proteiner behövde en fix struktur för att vara funktionella, och att dess struktur avgjorde funktionen. Detta ifrågasattes dock, när det konstaterades att en betydande del av alla proteiner faktiskt saknar väldefinierad struktur, men ändå är funktionella. Dessa kallas för oordnade proteiner och utmärker sig genom att vara flexibla och byta konformation ofta. Oordnade proteiner är involverade i många biologiska processer där deras brist på väldefinierad struktur faktiskt kan vara en fördel. Till exempel kan de lättare interagera med flera olika partners eftersom de är anpassningsbara, och därmed fungera bra för att reglera processer. När saker går snett med de oordnade proteinerna kan det dock uppstå sjukdomar. Alzheimers, Parkinsons, och vissa typer av cancer är alla exempel på sjukdomar som involverar oordnade proteiner. I vår saliv finns det också flertalet oordnade proteiner som hjälper till med att skydda tandemaljen och slemhinnor, samt att bekämpa virus, bakterier och svamp. Proteinet jag har jobbat mest med heter statherin och har som främsta funktion att binda kalciumsalter i saliven, så det finns lättillgängligt när emaljen måste byggas upp, men inte i så stora mängder att det bildas utfällningar. Genom att förstå hur oordnade proteiner fungerar kan vi förstå sjukdomsförlopp, hitta botemedel och hämta inspiration för utveckling av läkemedel.

Proteiner är uppbyggda som långa kedjor av aminosyror med olika karaktär. Det finns ca 20 olika aminosyror som naturligt ingår i proteiner, och beroende på vilka som ingår och i vilken ordning dessa är uppradade i proteinet, det vill säga vilken sekvens proteinet har, så får proteinet olika struktur och beteende. En av de största frågorna när det kommer till oordnade proteiner är hur den här relationen mellan sekvens, struktur och funktion faktiskt ser ut. För att få svar på det, måste vi studera många olika oordnade proteiner. Det är dock ganska svårt att bestämma struktur av oordnade proteiner, just eftersom de växlar mellan olika konformationer hela tiden och således vara utsträckta i ena stunden och mer kompakta i nästa stund. I de flesta experimentella tekniker som går att tillämpa på oordnade proteiner mäter man på jättemånga proteinmolekyler samtidigt och får ut ett medelvärde över tid. Man kan likna det vid att försöka få en bild av hur människor ser ut genom att ta ett långtidsexponerat foto på ett dansgolv, där de dansande människorna är proteinerna. Fotot kommer mest visa suddiga skuggor. Ett sätt att få en bättre bild av vad som försiggår är genom att använda sig av datorsimuleringar, vilket kan visa exakt hur varje protein ser ut i varje ögonblick, samtidigt som man kan beräkna medelvärden motsvarande den experimentella datan. För att kunna göra simuleringar behövs dock en modell. Modeller kan byggas upp på olika sätt, vilket illustreras i Figur 1. Ju mer detaljer som är med i modellen, desto mer detaljerad information kan fås ut, men det blir både svårare att tolka och mer krävande att simulera, i termer av datorresurser och tidsåtgång.



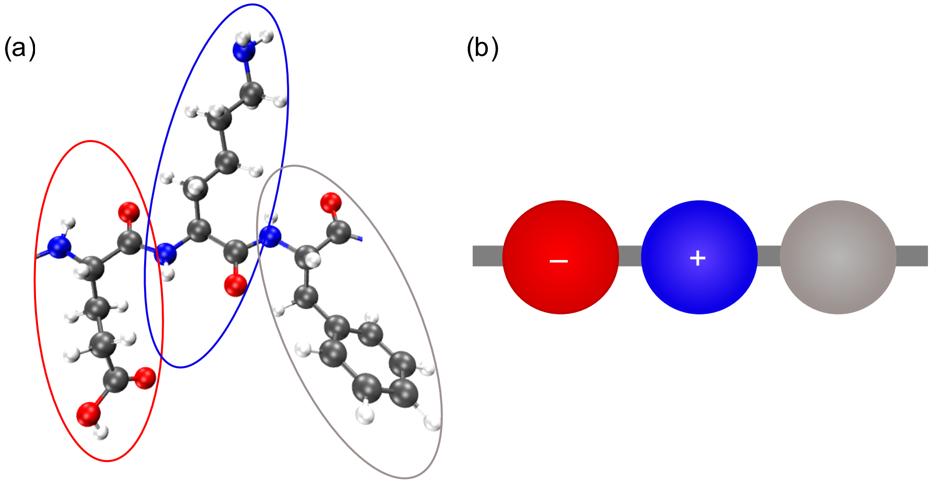
Figur 1: Olika modeller av en katt. Den till vänster är mest detaljerad. Modellerna till höger är grovkorniga och den längst till höger är mest grovkornig.

Beroende på vad vi har för forskningsfråga behöver vi därför ha olika modeller. För att fortsätta på exemplet med katten i Figur 1, så kan det vara viktigt att ha med svansen i en studie av hur katter kommunicerar. Om vi istället vill ta reda på hur många katter som får plats i ett rum räcker det dock med att se varje katt som en boll, vars storlek bestäms av hur stor katten är och hur mycket utrymme den vill ha. Men bara för att en modell innehåller mer detaljer betyder det inte att den ger bättre resultat. För att vara säkra på att modellerna stämmer och ger rätt resultat måste vi således ändå ha experimentella data att jämföra med.

I den här avhandlingen har jag främst haft två mål. Det första har varit att undersöka och vidareutveckla modeller för att beskriva oordnade proteiner, så att vi får fler verktyg för att studera denna typ av proteiner. Det andra har varit att undersöka sambandet mellan sekvens och struktur, framför allt hur fosforylering av proteiner påverkar strukturen. Fosforylering är en typ av reversibel ändring som kan göras på vissa aminosyror i ett protein, och som medför att aminosyran bland annat blir negativt laddad och får annan storlek. För att gå tillbaka till exemplet med katten, så kan vi likna det vid att sätta på katten en strumpa. Det kan påverka hur katten rör sig, och ha olika effekt beroende på vilken tass vi sätter den på, samt hur många tassar som får strumpor.

I mitt arbete har jag använt mig av två olika typer av modeller. Den första typen är en grovkornig modell, som beskriver ett protein som ett pärlhalsband. Varje pärla motsvarar en aminosyra, och har fått en laddning motsvarande den av aminosyran. Den andra typen är atomistisk, vilket innebär att alla atomer i alla aminosyror är representerade, så den är mycket mer detaljerad än den grovkorniga modellen, vilket visas i Figur 2. Den grovkorniga modellen visade sig kunna beskriva flertalet oordnade proteiner och ge en ökad förståelse för vad som kontrollerar proteinets struktur, det vill säga vilka konformationer det helst antar. En lite modifierad version av modellen kunde dessutom beskriva självassociering av statherin, det vill säga processen där flera proteinmolekyler går samman och bildar större kluster. Tillsammans med experimentella data kunde modellen användas för att avkoda vilka interaktioner som är viktiga i statherins självassociering. Den grovkorniga modellen visade sig dock överdriva hur kompakta proteiner som fosforylerats på många ställen är.

För att bättre förstå hur fosforylering påverkar proteiner behövdes en mer detaljerad modell



Figur 2: En bit av ett protein i en a) atomistisk modell och b) grovkornig modell. De färgade ovalerna visar vilka atomer som bakas samman till en pärla i den grovkorniga modellen.

än den grovkorniga, så därför använde jag två olika atomistiska modeller för att studera fosforylerade oordnade proteiner. Dessa modeller gav väldigt olika resultat, vilket visar vikten av att alltid jämföra med experiment. Den ena modellen visade sig kraftigt överskatta hur starka interaktionerna mellan fosforylerade och positivt laddade aminosyror är, vilket gjorde att proteinerna blev mer kompakta än vad experimentella metoder visade. Den andra modellen kunde kvalitativt fånga effekter av fosforylering som påvisats experimentellt och ge en detaljerad bild av vilka aminosyror som spelade roll och på vilket sätt. Detta visade att atomistiska simuleringar kan användas för att ge ökad förståelse av sambandet mellan sekvens och struktur, men att det är väldigt viktigt att fortsätta förbättra modeller.

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions**

C. Cragnell, **E. Rieloff**, M. Skepö
Journal of Molecular Biology, 2018, 430, 2478–2492.

II **Assessing the Intricate Balance of Intermolecular Interactions upon Self-association of Intrinsically Disordered Proteins**

E. Rieloff, M. D. Tully, M. Skepö
Journal of Molecular Biology, 2019, 431, 511–523.

III **Phosphorylation of a Disordered Peptide – Structural Effects and Force Field Inconsistencies**

E. Rieloff, M. Skepö
Journal of Chemical Theory and Computation, 2020, 16, 1924–1935.

IV **Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison**

E. Rieloff, M. Skepö
International Journal of Molecular Sciences (in press), 2021.

V **The Effect of Multisite Phosphorylation on the Conformational Properties of Intrinsically Disordered Proteins**

E. Rieloff, M. Skepö
Manuscript (submitted).

All papers are reproduced with permission of their respective publishers.

Publications not included in this thesis:

Determining R_g of IDPs from SAXS Data

E. Rieloff, M. Skepö
In: Kragelund B., Skriver K. (eds), Intrinsically Disordered Proteins. Methods in Molecular Biology, vol 2141. Humana, New York, NY

Author contributions

Paper i: Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

I performed the experiments and part of the simulations and analysis, took part in discussions and contributed to the writing of the paper.

Paper ii: Assessing the Intricate Balance of Intermolecular Interactions upon Self-association of Intrinsically Disordered Proteins

I planned the study together with my supervisor, performed the experiments and simulations and implemented cluster moves and analyses. I analysed the data with input from the co-authors, and wrote the manuscript with support from the co-authors.

Paper iii: Phosphorylation of a Disordered Peptide – Structural Effects and Force Field Inconsistencies

I planned the study together with my supervisor, performed the simulations, prepared the experimental samples, performed the circular dichroism spectroscopy experiments and analysed all the data. I wrote the manuscript with support from my supervisor and was responsible for the submission and revision process.

Paper iv: Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison

I planned the study together with my supervisor and performed the simulations and data analysis. I wrote the manuscript with support from my supervisor.

Paper v: The Effect of Multisite Phosphorylation on the Conformational Properties of Intrinsically Disordered Proteins

I planned the study together with my supervisor and performed all the experiments, simulations and data analysis. I wrote the manuscript with support from my supervisor.

List of abbreviations

A99	Amber ff99SB-ILDN + TIP4P-D
C36	CHARMM36m
CD	circular dichroism
CMC	critical micelle concentration
FCR	fraction of charged residues
FRET	fluorescence resonance energy transfer
IDP	intrinsically disordered protein
NCPR	net charge per residue
NMR	nuclear magnetic resonance
PBC	periodic boundary conditions
PCA	principal component analysis
PME	particle mesh Ewald
PTM	post-translational modification
R_g	radius of gyration
R_ee	end-to-end distance
SAXS	small-angle X-ray scattering

Acknowledgements

First I want to thank my supervisor *Marie* for all the support and guidance you have given me throughout the years. I also want to express my appreciation to all former and current group members and colleagues at the division, for forming a friendly environment, and providing good discussions and fun times at ”fika”. A special thanks to *Stephanie* and *Maria*, for all we have done together during these years. I am also thankful to *Carolina* for teaching me about experimental work with proteins and SAXS, and to *Mona*, *Eric*, and *Amanda* for reading and commenting on this thesis. Furthermore, I want to thank *my family* and my friend *Emil* for support. I feel endless gratitude towards *Max* for always being by my side and supporting me in all kinds of ways. Lastly, a huge thanks to *Ludvig*, for bringing me so much joy and showing me what is truly important in life.

Chapter 1

Introduction

For a long time, the structure–function paradigm dominated the view on proteins. According to this paradigm, protein function is critically dependent on a well-defined and folded three-dimensional structure, determined by sequence [1]. However, since the late 1990s, the field of intrinsically disordered proteins (IDPs) has rapidly evolved [2] and challenged this view. Despite being unfolded at physiological conditions, IDPs have proved to have important functions in our bodies [2–5] and are today recognised as an integral part of protein science. One of the main questions in this field is how sequence, structure, and function are related. Post-translational modification (PTM), such as phosphorylation, is a great example of how function can be regulated by modifications at the sequence level inducing structural changes.

Since IDPs lack well-defined structure they have proven more challenging to study experimentally than conventional proteins. Thus, computer simulations have emerged as a useful complement, to aid in the interpretation of experimental data and to access detailed information on the molecular level. Simulations are also useful for making predictions and investigations at conditions unattainable by experimental methods. However, to obtain successful results from computer simulations, accurate models are required. To this day, there is no model available that can describe everything, hence there is a wide range of specialised models. Simulations are also limited by the computational time and resources it takes to simulate a system, so different types of models are required for different research problems.

To evaluate models an important part is comparison with experimental data, hence, experiments and computer simulations are closely linked, and also in this thesis. The aims of this thesis have been: i) to contribute to the collection of possible tools to use for studying IDPs, by evaluation and further development of suitable models, and ii) to investigate the link between sequence and structure by studying conformational properties of IDPs in solution, with focus on phosphorylated IDPs.

Chapter 2

Background

This chapter describes IDPs and their biological relevance. The main part of my research has been focused around the saliva protein statherin, so it and its natural environment are given more focus.

2.1 Proteins

Proteins are biological macromolecules essential for life, as they provide a wide range of functions within organisms. Proteins are essentially polypeptides, since they are constructed as chains of amino acid residues connected by peptide bonds. Traditionally, the term protein is applied to long polypeptides consisting of 50 residues or more [6], while those shorter than that are referred to as polypeptides, or just peptides. Although there are many different amino acids, only roughly 20 are incorporated biosynthetically into proteins. These are referred to as *proteinogenic* amino acids. They all share the same basic structure, shown in Figure 2.1, consisting of an amino group ($-NH_2$), a carboxyl group ($-COOH$) and a side

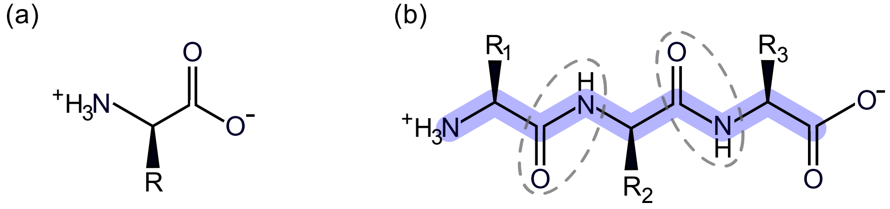


Figure 2.1: General structure of a) an amino acid and b) a tripeptide at pH 7, where R represents side groups. The backbone is highlighted in blue and the peptide bonds are shown within dashed ovals.

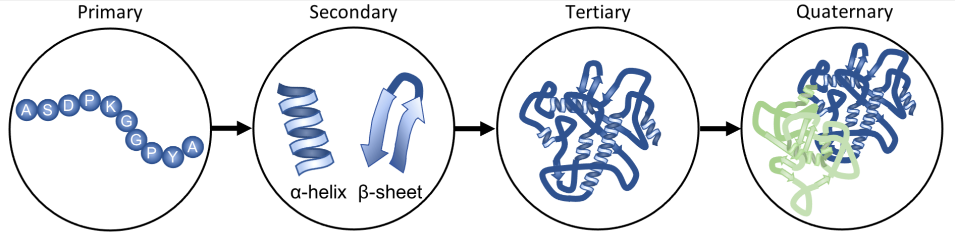


Figure 2.2: Illustration of the different levels of protein structure.

group ($-$R). At pH 7, which roughly corresponds to physiological pH, the amino group is protonised ($-\mathrm{NH_3}^+$) and the carboxyl group deprotonized ($-\mathrm{COO}^-$), making the amino acid *zwitterionic*. Depending on the characteristics of the side group, the amino acids can be classified as polar, hydrophobic, positively charged, or negatively charged.

The structure of a protein can be described at four different levels, as illustrated in Figure 2.2. The *primary structure* is the sequence of amino acid residues. Local parts of the chain can arrange into regular structures, referred to as *secondary structure*. The most common types of secondary structure are α-helix and β-sheet, which both form as a result of hydrogen bonds between protein backbone atoms [6]. 3_{10}- and π-helix are similar to α-helix, but differ in the hydrogen bond pattern, causing the pitch of the helix to be different. Turn is another rather common secondary structural element, which corresponds to a short segment in which the direction of the polypeptide chain is reversed. Another interesting type of secondary structure is the left-handed polyproline type II helix (PPII), which is a rather extended helix that actually lacks internal hydrogen bonds. Instead, it can be identified by the values of the backbone dihedral angles [7].

The protein can also fold into a well-defined three-dimensional shape, referred to as the *tertiary structure*. The major driving force behind folding is the hydrophobic interaction, trying to hide hydrophobic residues from the surrounding water [8]. In addition, a protein can consist of several different protein chains, each having a three-dimensional structure and making up a subunit of the complete protein. The arrangement of the subunits is called the *quaternary structure*.

2.2 Intrinsically disordered proteins

IDPs are characterized by a lack of well-defined tertiary structure under physiological conditions, which means that they are much more flexible than other proteins and interchange rapidly between many different conformations. Often can protein disorder be recognised already in the primary sequence. IDPs typically have a low sequence complexity and are generally enriched in charged and polar amino acids, with a low content of bulky hydrophobic amino acids [9, 10].

When IDPs and intrinsically disordered regions first were discovered, they were regarded as non-functional and of no importance, due to the belief that protein function was strongly coupled to the three-dimensional structure. Since then, it has been shown that intrinsic disorder is actually wide-spread in nature. At least 10% of eukaryotic proteins are intrinsically disordered, while even more proteins contain long disordered regions [11–14]. In addition, it has been established that IDPs are involved in many important biological processes, such as regulation, signalling, and recognition, where intrinsic disorder can actually be crucial for the function [3–5, 13, 15–17]. Some advantages of disorder are that it enables interactions of high specificity coupled with low affinity, multiple binding partners, faster association/disassociation rates, and larger interaction surfaces [4]. Furthermore, many IDPs have been shown to have folding induced upon binding to interaction partners [2, 4, 18]. Due to the immense biological functions of IDPs, there is no surprise that they are also associated with pathological conditions, for example Alzheimer's disease, Parkinson's disease, diabetes, and several types of cancer [19, 20].

2.2.1 Classification of IDPs

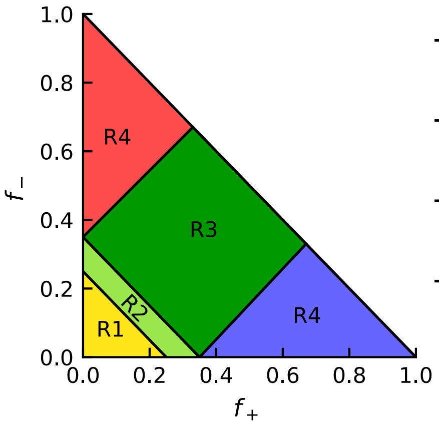
IDPs are a rather heterogeneous group, including less or more compact proteins with different degrees of secondary and tertiary structure [21, 22]. The amino acid composition and charge distribution have been shown to be important for the conformational properties of IDPs, such that they can be used to define conformational classes. From the fraction of positively and negatively charged residues, f_+ and f_-, the fraction of charged residues (FCR) and net charge per residue (NCPR) are defined according to

$$\mathrm{FCR} = f_+ + f_- \tag{2.1}$$

$$\mathrm{NCPR} = |f_+ - f_-|. \tag{2.2}$$

Based on these quantities, Das et al. have introduced a diagram-of-state with four different conformational classes called R1–R4 [23], shown in Figure 2.3. The R1 class consists of globules, while the R3 class are made up by coils and hairpins. The R2 class is an intermediate region, such that IDPs in this class usually adopt both coil and more globule-like conformations. The IDPs in the R4 class are either strongly positively or negatively charged, and behave as semi-flexible rods or coils.

Polymers consisting of positively or negatively charged subunits are called *polyelectrolytes*, while polymers containing subunits of mixed charges are called *polyampholytes*. They can be either weak or strong, depending on their FCR. Applying this terminology to IDPs, weak polyampholytes and polyelectrolytes are found in the R1 class, strong polyampholytes in the



Class	FCR	NCPR	Conformation
R1	<0.25	< 0.25	
R2	0.25– 0.35	≤0.35	
R3	>0.35	≤0.35	
R4	>0.35	>0.35	

Figure 2.3: Diagram-of-states showing conformational classes of IDPs based on the fraction of positively (f_+) and negatively (f_-) charged residues, fraction of charged residues (FCR), and net charge per residue (NCPR), as introduced by Das et al. [23]. R1: globules, R2: mix of globules and coils, R3: coils or hairpins, R4: semi-flexible rods or coils.

R3 class, and strong polyelectrolytes in the R4 class. This classification scheme to predict the conformational class of an IDP is valid for IPDs consisting of at least 30 residues, having low hydrophobicity and low proline content. A high proline content is expected to give more extended conformations than the diagram-of-states predicts.

For the IDPs in the R3 class, the distribution of charges throughout the sequence also determines what conformations are adopted. The distribution of charges can be described using the parameter κ, loosely described as a parameter accounting for charge mixing. κ adopts a value between zero and one, where the maximum value corresponds to the sequence with the largest possible segregation of opposite charges for the given composition. IDPs having a low κ are expected to behave more as self-avoiding random walks, while IDPs with a high κ are more likely to adopt hair-pin like conformations. κ can also be useful for predicting the influence of salt concentration, since IDPs with high κ usually show larger conformational changes upon changes in ionic strength [24].

2.3 Phosphorylation

A common regulatory strategy employed by cells is PTM, in which a protein is chemically modified after synthesis by for example the addition of a modifying group. One of the most abundant PTM is phosphorylation, in which a phosphoryl group is attached to a residue, most commonly serine or threonine. Phosphorylation is a reversible process, and especially prevalent among IDPs and disordered regions [4, 25, 26]. As seen in Figure 2.4,

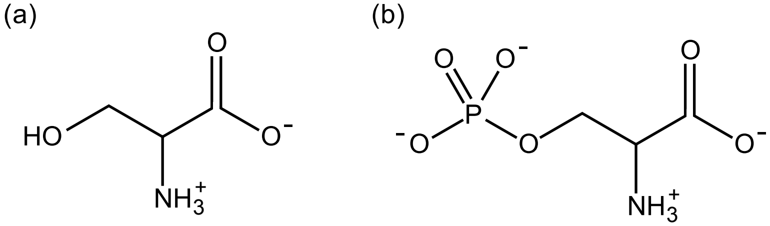


Figure 2.4: The structure of a) serine and b) phosphoserine at physiological pH.

phosphorylation increases the bulkiness of the residue and introduces two additional negative charges at physiological pH, which can greatly influence the electrostatic interactions within a protein or with a binding partner. It has been established that phosphorylation can induce changes in both overall conformation and secondary structure, as well as affect the dynamics and interactions with binding partners [27]. As a consequence, abnormal phosphorylation can be pathological; for example, Alzheimer's disease is associated with hyperphosphorylation of the neuroprotein tau [28]. In the disordered milk proteins caseins and saliva protein statherin, phosphorylated residues are of direct importance for the functionality, by enabling sequestration of calcium [29] and increasing binding to the tooth surface [30, 31].

2.4 Saliva

Saliva is a complex fluid of great importance to our oral health, even though it consists of 99.5% water. The rest involves inorganic components such as sodium, potassium, calcium, and chloride, and organic components such as proteins, lipids, and carbohydrates. Saliva aids speaking and swallowing through lubrication of the oral tissues, helps with digestion, provides protection for the teeth, and is a first line of defence against bacteria, viruses, and fungii [32]. Many of the protective functions of saliva are attributed to proteins, as presented in Figure 2.5. Note that several of these proteins are in fact intrinsically disordered and multi-functional. Many of the proteins are part of the acquired enamel pellicle, which is a thin protein-rich film that forms on the tooth surface. The pellicle protects against acid degradation, provides lubrication that protects the teeth from abrasion and attrition, and also serves as a layer to which bacteria can adhere [33, 34].

The composition, and hence the ionic strength and pH of saliva, varies with a lot of different factors, for example time of day and food intake. The saliva production can also be affected by diseases and medication [33].

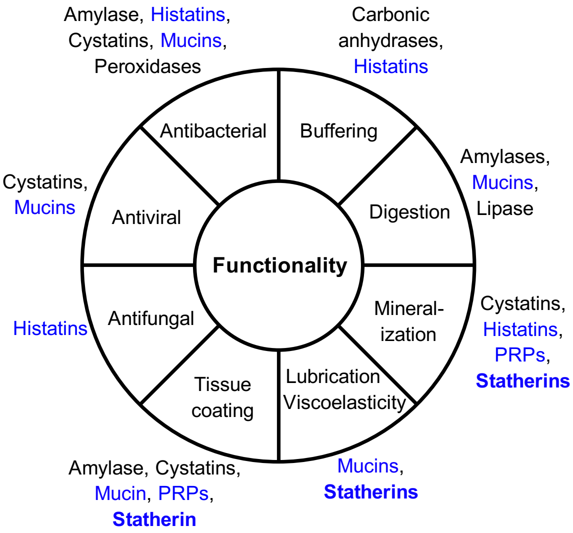


Figure 2.5: Proteins responsible for functionality of saliva, where intrinsically disordered proteins are marked in blue. The figure is adapted from Levine [35].

2.5 Statherin

Statherin is one of the intrinsically disordered salivary proteins that is part of the aquired enamel pellicle. The main function of statherin is to prevent spontaneous precipitation of calcium phosphate salts in saliva, in order to maintain a supersaturated environment [36, 37], which helps with remineralisation after dental erosion [38]. In addition, statherin has also been shown to have lubricative properties [39] and promote adhesion of certain bacteria that are associated with cemental caries and gum disease [40–42].

Statherin is a rather small protein, only 43 amino acids long with a molecular weight of 5.38 kDa, which makes it suitable for modelling. It has a distinct charge distribution, evident in the primary sequence in Figure 2.6, where nine out of ten charged residues are located among the first 13 residues in the N-terminal part. This N-terminal part, including the acidic motif with two phosphorylated serines, has been shown to be of extra importance for the ability of statherin to adsorb to the tooth enamel and prevent crystal growth [30]. Overall, the hydrophobicity is rather low (based on the hydropathy values in the Kyte-Doolittle scale [44]), which is typical for IDPs. However, region 15–43 is rich in prolines and glutamines, which allow for weak association to many other proteins [45], and contain seven tyrosines, whose aromatic side-chains have been established to be of importance for liquid-liquid phase separation [46, 47]. Statherin self-associates upon increased protein concentration [48], such that several protein chains merge to a larger complex. Self-association is further described in the following section.

```
+DSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF-
```

Figure 2.6: The primary sequence of Statherin [43]. Amino acids that have a negatively charged side chain at pH 8 are marked in red, and those with a positively charged side chain are marked in blue. The phosphorylated serines (marked in dark red) have a charge of -2*e* each at pH 8.

2.6 Self-association

Self-association is the spontaneous formation of larger structures from smaller constituents. A typical example of self-association is the micelle formation of surfactants. Surfactants usually consist of a hydrophobic tail and a polar head-group, which means that they are *amphiphilic*. Driven by the hydrophobic interaction (see section 3.9) the surfactants arrange into spherical structures called micelles, hiding the hydrophobic tails in the interior, as shown in Figure 2.7. This only happens above a certain surfactant concentration, named the *critical micelle concentration* (CMC).

Self-association is governed by intermolecular interactions, such as van der Waals interactions, hydrogen bonding, hydrophobic interaction, and screened electrostatic interactions, which are further described in chapter 3. Since these interactions are generally weak, at least compared to covalent bonds, the self-association process is highly affected by solution conditions such as pH and ionic strength. Both the interactions between and within self-assembled structures are affected by changes in the solution conditions, therefore the size and shape of the self-assembled complexes can be modified [49].

Large molecules such as amphiphilic block-copolymers can also form micelles, however, due to their much larger size and sometimes more pronounced amphiphilic nature, the behaviour can differ from surfactants. Proteins can also self-associate, which the intrinsically disordered milk protein β-casein is a good example of. The C-terminal part of β-casein contains many hydrophobic residues, while the N-terminal part has several phosphorylated residues that contributes to a net charge, giving the protein chain an amphiphilic structure. Many studies, only a few mentioned here, have been devoted to the β-casein micelle form-

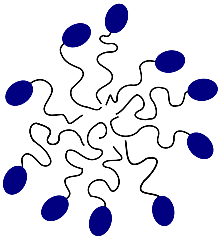


Figure 2.7: A schematic illustration of a micelle formed of surfactants having polar head-groups and hydrophobic tails.

ation and have shown that the micelle size and shape, as well as CMC are sensitive to the solution conditions such as temperature, pH and protein concentration [50–54].

Chapter 3

Intermolecular interactions

Studying proteins from a chemical point of view, we distinguish between two classes of interactions: i) covalent bonds that keep the atoms together in molecules, and ii) non-covalent intermolecular interactions. Although the term *intermolecular* literarily translates to existing or occurring *between* molecules, the interactions also act between different parts of molecules. The intermolecular interactions are generally weak compared to covalent bonds, but are highly important as they account for how proteins behave, for example how they fold and bind to other molecules. The intermolecular interactions that will be described in this chapter can be classified as short-ranged or long-ranged, depending on their distance dependence. The van der Waals interaction, having a $1/r^6$-dependence, is a typical example of a short-ranged interaction, while the Coulomb interaction acting between charged species is considered long-ranged, due to its $1/r$-dependence. The decay of potentials with different distance dependence is shown in Figure 3.1. This chapter is mostly based on the book by Israelachvili [49], which is referred to for a more thorough description.

3.1 Charge–charge interaction

The electrostatic force, F, between two atoms with charges Q_i and Q_j, separated by a distance r, is described by the Coulomb law

$$F(r) = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r^2}, \tag{3.1}$$

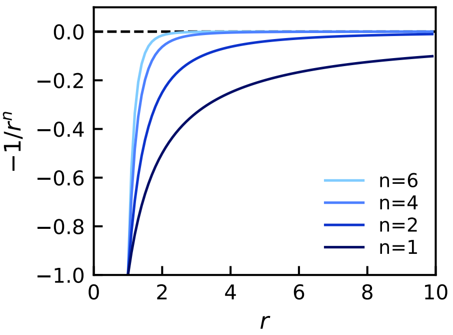


Figure 3.1: Illustration of the decay of potentials with different distance dependence.

where ε_0 is the vacuum permittivity and ε_r is the relative permittivity of the surrounding medium. The interaction free energy, $w(r)$, between the two charges is given by

$$w(r) = \int_0^\infty -F(r)\mathrm{d}r = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r}. \tag{3.2}$$

The interaction is long-ranged, but if the charges are surrounded by ions, as in an aqueous salt solution, the interaction is screened, which reduces the range of the interaction. According to the Debye–Hückel theory, a screened Coulomb potential can be expressed as

$$V(r) = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}} \frac{1}{r} \exp(-\kappa r), \tag{3.3}$$

where $V(r)$ is the potential energy and κ^{-1} is the Debye length, defined by

$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0\varepsilon_\mathrm{r} kT}{2N_\mathrm{A} e^2 I}}, \tag{3.4}$$

where k is the Boltzmann constant, T is the temperature, N_A the Avogadro constant, e the elementary charge, and I refers to the ionic strength, defined as

$$I = \frac{1}{2}\sum_{i=1}^{n} c_i Z_i^2. \tag{3.5}$$

Here, n is the number of different ion species, and c_i is the concentration of ion i with charge number Z_i.

3.2 Charge–dipole interaction

Most molecules have no net charge; however, they often possess an electric dipole, caused by an asymmetric distribution of electrons in the molecule. The dipole moment is defined as

$$\mu = q\mathbf{l}, \tag{3.6}$$

where $\mathbf{l}$ is the distance vector between the two charges $-q$ and $+q$. When a charge and a dipole interact at a distance $r >> l$, the potential energy is given by

$$V(r, \theta) = -\frac{Q\mu\cos\theta}{4\pi\varepsilon_0\varepsilon_\mathrm{r}}\frac{1}{r^2}, \tag{3.7}$$

where the polar angle, θ, is the angle between the distance vector and the dipole (see Figure 3.2a). If the charge is positive, maximum attraction occurs when the dipole points away from the charge ($\theta = 0°$). At large separation or in a medium with high relative permittivity, the angle dependence of the interaction can fall below the thermal energy kT, which allows the dipole to rotate more or less freely. However, conformations allowing for attractive interactions will still be more favourable, so the angle-averaged potential will not be zero. The interaction free energy between a freely rotating dipole and a charge is given by

$$w(r) \approx -\frac{Q^2\mu^2}{6(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2 kT}\frac{1}{r^4} \text{ for } kT > \frac{Q\mu}{4\pi\varepsilon_0\varepsilon_\mathrm{r}r^2}. \tag{3.8}$$

Note that this changes the distance dependence of the potential, making it more short-ranged.

3.3 Dipole–dipole interaction

The interaction energy between two stationary dipoles i and j can be described by the following potential

$$V(r, \theta_i, \theta_j, \phi) = -\frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}}\frac{1}{r^3}(2\cos\theta_i\cos\theta_j - \sin\theta_i\sin\theta_j\cos\phi), \tag{3.9}$$

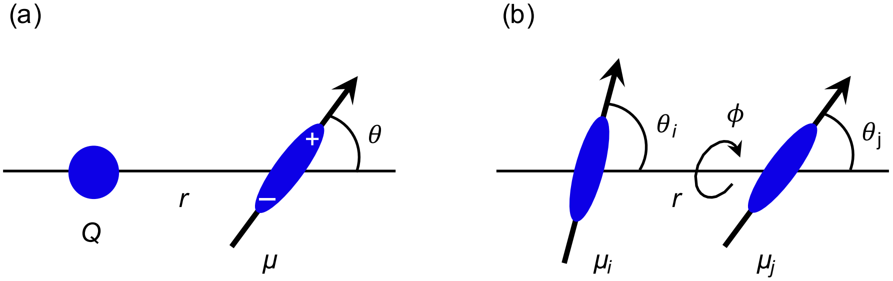


Figure 3.2: Schematic representation of the (a) charge–dipole and (b) dipole–dipole interaction, where r is the distance between the interacting species, θ is the polar angle and ϕ the azimuthal angle.

where ϕ is the azimuthal angle between the dipoles (see Figure 3.2b). Also in this case can the dipoles rotate, so the angle-averaged interaction free energy is

$$w(r) = -\frac{\mu_i^2 \mu_j^2}{3(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2 kT} \frac{1}{r^6} \text{ for } kT > \frac{\mu_i \mu_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r} r^3}. \tag{3.10}$$

This interaction is usually referred to as the *Keesom interaction* and is a part of the total van der Waals interaction described in section 3.6.

3.4 Charge–induced dipole interaction

All molecules and atoms, even non-polar ones, are polarised by an external electric field, which means that the electron cloud in the molecule is displaced. Hence, the electric field exhibited by a charge will induce a dipole moment in a non-polar molecule. The potential between the charge and the induced dipole is expressed as

$$V(r) = -\frac{-Q^2\alpha}{2(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2} \frac{1}{r^4}, \tag{3.11}$$

where α is the polarisability of the molecule.

3.5 Dipole–induced dipole interaction

Similarly to the charge–induced dipole interaction, a non-polar molecule can gain an induced dipole moment in the field from a permanent dipole. The interaction is described by the following potential,

$$V(r) = -\frac{\mu^2\alpha}{(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2} \frac{1}{r^6}. \tag{3.12}$$

Notice that this potential is already angle-averaged, since the interaction normally is not strong enough to mutually orient the molecules. This interaction is usually referred to as the *Debye interaction* and is a part of the total van der Waals interaction due to the $1/r^6$-dependence.

3.6 Van der Waals interaction

The total van der Waals interaction includes three different types of interactions, which all have a $1/r^6$-dependence: Keesom, Debye and London (dispersion), of which Keesom and Debye have been described above (section 3.3 and 3.5). The Keesom interaction is only present between permanent dipoles and the Debye interaction when one of the molecules is a permanent dipole. The last interaction, the *London dispersion interaction* is however present between all types of molecules. It is of quantum mechanical origin, although we can think of it in a simpler manner. For a non-polar atom (or molecule) the time averaged dipole moment is zero, although at any instant it exists a finite dipole moment caused by an uneven electron distribution around the nucleus. This instantaneous dipole generates an electric field that induces a dipole in another nearby atom (or molecule), leading to an attractive interaction.

3.7 Hydrogen bond

In the previous chapter hydrogen bonds where mentioned in the context of protein secondary structure. A hydrogen bond can occur between a highly electronegative atom, such as nitrogen, oxygen or fluorine, and a hydrogen covalently bonded to another such electronegative atom. It is of predominantly electrostatic origin and can be seen as an especially strong dipole–dipole interaction. Unlike normal dipole–dipole interactions it is fairly directional and can be described by a $1/r^2$-dependence, similar to the charge–dipole interaction.

3.8 Exchange repulsion (excluded volume)

At very small interatomic distances, when electron clouds overlap, a strong repulsive interaction of quantum mechanical origin occurs, which limits how close two atoms can come. The repulsion increases steeply with decreased distance and is therefore often modelled with a hard sphere potential which goes directly from zero to infinity, or with a soft core potential of $1/r^{12}$-dependence.

3.9 Hydrophobic interaction

Water is a special solvent due to the possibility to form many hydrogen bonds, which makes the water–water interaction strong. Therefore, the water molecules much rather interact with other water molecules than non-polar molecules. For small non-polar molecules the water can arrange around the non-polar molecule in such a way that no hydrogen bonds are broken. However, this arrangement is more ordered and therefore comes at an entropic cost, which makes it more favourable to separate the non-polar molecules from the water molecules. For large non-polar molecules it is not possible to retain hydrogen bonds, which instead leads to an energy driven separation. Therefore, the cause of separation between water and non-polar molecules can be both mostly entropic or mostly energetic, however, the net result can always be seen as an effective attraction between non-polar molecules, called a hydrophobic interaction [55].

3.10 Conformational entropy

When a flexible polymer, for example an IDP, approaches a surface or other polymers, restrictions are enforced on the available conformations, which leads to a decrease in conformational entropy. If the restrictions are large enough, the result will be an effective repulsion of entropic origin.

Chapter 4

Statistical thermodynamics

Statistical mechanics provides a connection between macroscopic properties, such as temperature and pressure, and microscopic properties related to the molecules and their interactions. The aim is to provide means to both predict macroscopic phenomenas and understand them on a molecular level. Statistical mechanics applied for explaining thermodynamics is usually referred to as statistical thermodynamics. Here I will provide a brief introduction to the key concepts, while a more in-depth description can be found in for example the book by Hill [56].

A central concept in statistical mechanics is *ensembles*. An ensemble is an imaginary collection of a very large number of systems, each being equal at a thermodynamic (macroscopic) level, but differing on the microscopic level. Ensembles can be classified according to the macroscopic system that they represent, as outlined below.

Microcanonical (NVE) **ensemble**: represents an isolated system in which the number of particles (N), the volume (V) and the energy (E) are constant. Hence, the systems in the ensemble all have the same N, V, and E, and share the same environment, however, they correspond to different microstates.

Canonical (NVT) **ensemble:** corresponds to a closed and isothermal system, by having constant number of particles, volume, and temperature (T).

Grand canonical ensemble (μVT)**:** represents an open isothermal system, in which the chemical potential (μ), the volume, and the temperature are kept constant.

Isothermal-isobaric ensemble (NpT)**:** has constant number of particles, pressure (p), and temperature.

When an experimental measurement is performed, a time average is taken over the observable of interest. If we instead want to calculate the observable from molecular properties, we would need to deal with both a large number of molecules and the requirement to observe them for a sufficiently long time to smear out molecular fluctuations. In practice this would be extremely complicated, however, a different approach is possible due to the *first postulate of statistical mechanics*: a (long) time average of a mechanical variable in a thermodynamic system is equal to the ensemble average of the variable in the limit of an infinitely large ensemble, provided that the ensemble replicate the thermodynamic state and environment. Stated differently, this postulate says that instead of using a time average, we can obtain the same result by performing an ensemble average, given that the ensemble is sufficiently large. This is valid for all ensembles and provides the basis for molecular simulations. There is also a *second postulate of statistical mechanics* which states that for an infinitely large ensemble representing an isolated thermodynamic system, the systems of the ensemble are distributed uniformly over the possible states consistent with the specified values of N, V and E. This postulate is also referred to as the *principle of equal a priori probabilities*, as it says that in the microcanonical ensemble, all microscopic states are equally probable.

In the canonical ensemble, the probability to find the system in a particular energy state E_i is

$$P_i(N, V, T) = \frac{\exp[-E_i(N, V)/kT]}{Q(N, V, T)}, \tag{4.1}$$

where Q is the canonical partition function, given by

$$Q(N, V, T) = \sum_i \exp[-E_i(N, V)/kT], \tag{4.2}$$

where $\exp[-E_i(N, V)/kT]$ is known as the Boltzmann weight. The partition function describes the equilibrium statistical properties of the system and can be used to express the Helmholtz free energy, A, as

$$A = -kT \ln Q. \tag{4.3}$$

The Helmholtz free energy is the characteristic function for the canonical ensemble and can be used to derive other thermodynamic variables, such as the entropy, pressure and total energy.

Here the partition function has been introduced in a quantum mechanical formulation with discrete energy states. However, many simulation methods are based on classical mechanics, in which the microstates are so close in energy that they are approximated as a continuum. In a classical treatment the canonical partition function becomes

$$Q_{\text{class}} = \frac{1}{N!h^{3N}} \int \exp[-H(\mathbf{p}^N, \mathbf{r}^N)/kT] \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N, \tag{4.4}$$

where h is Planck's constant and the integration is performed over all momenta $\mathbf{p}^N$ and all coordinates $\mathbf{r}^N$ for all N particles. $H(\mathbf{p}^N, \mathbf{r}^N)$ is the Hamiltonian of the system, having one kinetic energy part (dependent on the temperature) and one potential energy part (dependent on the interactions). The kinetic part can be integrated directly, simplifying the partition function to

$$Q_{\text{class}} = \frac{Z_N}{N!\Lambda^{3N}}, \tag{4.5}$$

where

$$Z_N = \int_V \exp[-U_{\text{pot}}(\mathbf{r}^N)/kT]\mathrm{d}\mathbf{r}^N \tag{4.6}$$

is the configurational integral calculated from the potential energy, U_{pot}, and

$$\Lambda = \frac{h}{(2\pi mkT)^{1/2}} \tag{4.7}$$

is the de Broglie wavelength, where m is the mass. If we know the configurational integral, we can calculate the ensemble average of an observable X, according to

$$\langle X(\mathbf{r}^N)\rangle = \frac{\int_V X(\mathbf{r}^N)\exp[-U_{\text{pot}}(\mathbf{r}^N)/kT]\mathrm{d}\mathbf{r}^N}{Z_N}. \tag{4.8}$$

However, solving the integrals is normally a rather challenging problem that requires numerical solution tools, such as the Monte Carlo method that will be discussed in chapter 6.

Chapter 5

Simulation models

A model is a representation of reality and can be constructed with varying degree of detail. When constructing or choosing a model, it is important to consider the properties of interest. The model should include enough detail to be able to accurately describe the properties of interest. Including excessive detail makes the model harder to interpret and increases the computational cost, which can limit the accessible time scale or system size. Hence, different scientific problems requires different models. In this thesis, two different types of models have been used to study IDPs, specifically a coarse-grained model representing each amino acid as a hard sphere, and an atomistic model including all atoms in the system, see Figure 5.1.

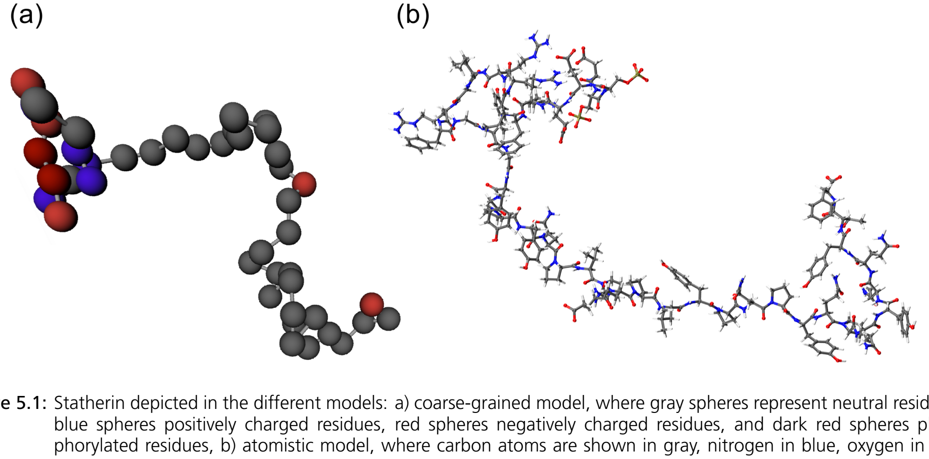


Figure 5.1: Statherin depicted in the different models: a) coarse-grained model, where gray spheres represent neutral residues, blue spheres positively charged residues, red spheres negatively charged residues, and dark red spheres phosphorylated residues, b) atomistic model, where carbon atoms are shown in gray, nitrogen in blue, oxygen in red, hydrogen in white, and phosphorus in tan.

5.1 The coarse-grained model

The coarse-grained model is a bead-necklace model based on the primitive model, in which each amino acid is described as a hard sphere (bead), connected by harmonic bonds. The N- and C-termini are modelled explicitly as charged spheres in each end of the protein chain, so the full length corresponds to the number of amino acids plus two. Each bead has a fixed point charge of $+1e$, 0, $-1e$, or $-2e$, corresponding to the state of the amino acid side chain at the desired pH. The counterions are included explicitly, while the solvent (water) and salt is treated implicitly. The model, as used in Paper 1, was parameterised by Cragnell et al. for the saliva IDP histatin 5 [57].

The model contains contributions from excluded volume, electrostatic interactions, and a short-ranged attraction mimicking van der Waals-interactions. The total potential energy is divided into bonded and non-bonded interactions, according to

$$U_{\text{tot}} = U_{\text{bond}} + U_{\text{non-bond}} = U_{\text{bond}} + U_{\text{hs}} + U_{\text{el}} + U_{\text{short}}, \tag{5.1}$$

where U_{hs} is a hard-sphere potential, U_{el} the electrostatic potential, and U_{short} a short-ranged attraction. The non-bonded energy is assumed pairwise additive, according to

$$U_{\text{non-bond}} = \sum_{i<j} u_{ij}(r_{ij}), \tag{5.2}$$

where u_{ij} is the interaction between two particles, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the center-to-center distance between the two particles, and $\mathbf{r}$ refers to the coordinate vector.

A harmonic bond represents the bonded interaction,

$$U_{\text{bond}} = \sum_{i=1}^{N-1} \frac{k_{\text{bond}}}{2}(r_{i,i+1} - r_0)^2. \tag{5.3}$$

Here, N denotes the number of beads in the protein, k_{bond} is the force constant having a value of 0.4 N/m, and $r_{i,i+1}$ is the center-to-center distance between two connected beads, with the equilibrium separation r_0 = 4.1 Å.

The excluded volume is accounted for by a hard sphere potential,

$$U_{\text{hs}} = \sum_{i<j} u_{ij}^{\text{hs}}(r_{ij}), \tag{5.4}$$

where the summation extends over all beads and ions. Here, u_{ij}^{hs} represents the hard sphere potential between two particles, according to

$$u_{ij}^{\text{hs}}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j \\ \infty, & r_{ij} < R_i + R_j \end{cases}, \tag{5.5}$$

where R_i and R_j denote the radii of the particles (2 Å). The electrostatic potential energy is given by an extended Debye–Hückel potential,

$$U_{\mathrm{el}} = \sum_{i<j} u_{ij}^{\mathrm{el}}(r_{ij}) = \sum_{i<j} \frac{Z_i Z_j e^2}{4\pi\varepsilon_0\varepsilon_{\mathrm{r}}} \frac{\exp[-\kappa(r_{ij} - (R_i + R_j))]}{(1+\kappa R_i)(1+\kappa R_j)} \frac{1}{r_{ij}}. \tag{5.6}$$

Hence, the salt in the system is treated implicitly as a screening of the electrostatic interactions.

The short-ranged attractive interaction is expressed as

$$U_{\mathrm{short}} = -\sum_{i<j} \frac{\varepsilon_{\mathrm{short}}}{r_{ij}^6}, \tag{5.7}$$

where summation extends over all beads. Here, $\varepsilon_{\mathrm{short}}$ reflects an average amino acid polarisability and sets the strength of the attraction. In this model $\varepsilon_{\mathrm{short}}$ is $0.6 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of $0.6\,kT$ at closest contact.

In Paper II, an additional short-ranged interaction is included in the model, to make the protein chains associate. This mimicks a hydrophobic interaction, which is applied between all neutral amino acids, according to

$$U_{\text{h-phob}} = -\sum_{\text{neutral}} \frac{\varepsilon_{\text{h-phob}}}{r_{ij}^6}, \tag{5.8}$$

where $\varepsilon_{\text{h-phob}}$ is $1.32 \cdot 10^4$ kJ Å/mol. This corresponds to an attraction of $1.32\,kT$ at closest contact. The value of $\varepsilon_{\text{hphob}}$ was set by comparing the average association number with experimental results obtained by small-angle X-ray scattering (SAXS).

5.2 The atomistic model

In the atomistic model, distributed in the GROMACS simulation package [58–62], each atom in the system is included, hence, also solvent molecules and ions are modelled explicitly. The total potential energy consists of bonded and non-bonded interactions, according to

$$U_{\mathrm{tot}} = \underbrace{U_{\mathrm{bond}} + U_{\mathrm{angle}} + U_{\mathrm{d}} + U_{\mathrm{id}}}_{\text{bonded}} + \underbrace{U_{\mathrm{LJ}} + U_{\mathrm{el}}}_{\text{non-bonded}}. \tag{5.9}$$

The bonded potentials act on covalently bonded atoms and each of the interaction potentials are summed over the atoms involved in the interaction. The first bonded term is a harmonic potential representing bond stretching,

$$U_{\mathrm{bond}} = \sum_b \frac{1}{2} k_{ij}^{\mathrm{b}} \left(r_{ij} - r_{ij}^0\right)^2, \tag{5.10}$$

where k_{ij}^{b} is a force constant, r_{ij} the distance between two bonded atoms i and j, and r_{ij}^{0} the equilibrium bond length. The second term is the bond angle vibration,

$$U_{\text{angle}} = \sum_{\theta} \frac{1}{2} k_{ij}^{\theta} \left(\theta_{ijk} - \theta_{ijk}^{0} \right)^{2}, \tag{5.11}$$

in which k_{ij}^{θ} is a force constant, and θ_{ijk} the angle between the three atoms i-j-k, having the equilibrium angle θ_{ijk}^{0}. The third and fourth term are torsion potentials related to dihedral angles, i.e. angles between two intersecting planes, controlling the rotation of a bond around its own longitudinal axis. Here, the proper dihedral angle is defined according to the IUPAC/IUB convention [63], as the angle ϕ_{ijkl} between the ijk and jkl planes, with zero corresponding to the cis conformation (atoms i and l on the same side). The proper dihedral angle potential is given by a sinusoidal function with periodicity n and phase ϕ_{s}:

$$U_{\mathrm{d}} = \sum_{\phi} k_{\phi} \left[1 + \cos(n\phi_{ijkl} - \phi_{\mathrm{s}}) \right], \tag{5.12}$$

where k_{ϕ} is a force constant. Unlike for the proper dihedrals, the atoms defining an improper dihedral do not need to be linearly connected. The improper dihedrals are used to keep planar groups (e.g. aromatic rings) planar, and maintain chirality. The improper dihedral angle potential is a harmonic potential,

$$U_{\mathrm{id}} = \sum_{\xi} \frac{1}{2} k_{\xi} \left(\xi_{ijkl} - \xi_{0} \right)^{2}, \tag{5.13}$$

where k_{ξ} is the force constant and ξ_{ijkl} the angle between the planes having an equilibrium dihedral angle ξ_{0}. The bonded interactions are illustrated in Figure 5.2.

Regarding the non-bonded interaction potentials, both are assumed pairwise additive. The Lennard-Jones potential,

$$U_{\mathrm{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5.14}$$

represents steric repulsion and an attractive dispersion interaction. Here, ϵ_{ij} is the depth of the potential well, and σ_{ij} corresponds to the finite distance at which the potential becomes zero. For the force fields used in this work, the Lorentz-Berthelot rules are used to calculate ϵ_{ij} and σ_{ij}, according to

$$\begin{aligned} \epsilon_{ij} &= (\epsilon_{ii}\epsilon_{jj})^{1/2}, \\ \sigma_{ij} &= \frac{\sigma_{ii} + \sigma_{jj}}{2}. \end{aligned} \tag{5.15}$$

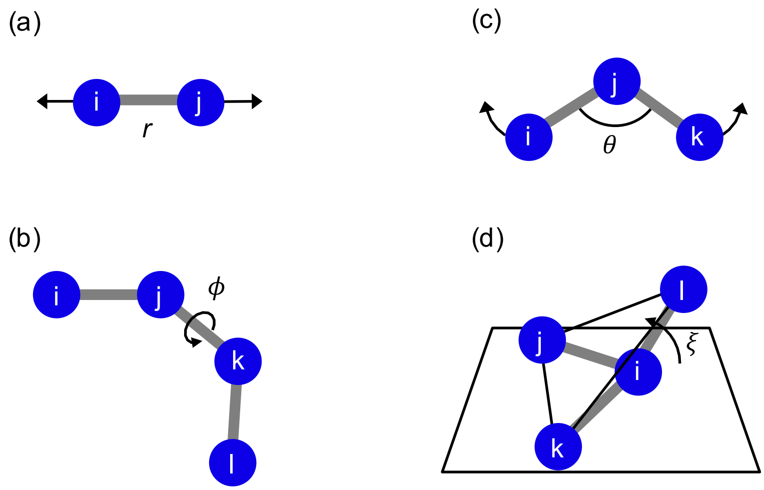


Figure 5.2: Schematic representation of the bonded interactions included in the atomistic model: a) bond stretching, b) bond angle vibration, c) proper dihedral torsion, and d) improper dihedral torsion.

The electrostatic interactions are represented by the Coulomb interaction,

$$U_{\text{el}} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{\text{r}} r_{ij}}, \tag{5.16}$$

where q_i and q_j are the charges of particle i and j, respectively.

5.2.1 Explicit water models

As previously mentioned, the atomistic simulations include the solvent, i.e. water, explicitly. The reason for this, is that the solvent itself and solvent–biomolecule interactions can have critical influence for biomolecules immersed in solvent. In fact, IDPs have been shown to be especially sensitive to how the water is represented, due to the extended conformations often adopted significantly exposing the protein to solvent [64–66].

There are many different explicit water models available, and due to the large number of water molecules needed to simulate a biomolecular system, the level of complexity of the water model not only influences the accuracy, but also the computational time. Among the most widely used water models today are the rigid point-charge water models with pairwise additive interactions. Due to having a fixed geometry of the water molecule, only non-bonded interactions (Coulomb and Lennard-Jones interactions) are included explicitly, which reduces the required computational effort [67]. The water models can be further dived into classes based on the number of interaction sites they contain. As shown in

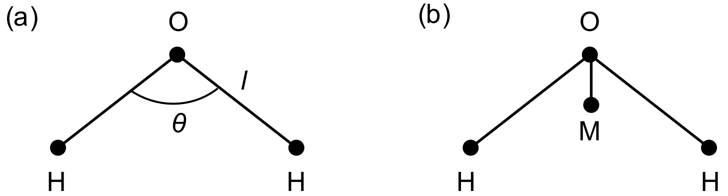


Figure 5.3: Illustration of a a) three-site and b) four-site water model, with the bond length l and bond angle θ. M represents a dummy atom where the oxygen charge is located.

Figure 5.3, three-site models have three sites, one for each atom in the molecule. In four-site models the oxygen charge is displaced to a fourth site M, while the Lennard-Jones term remains on the oxygen. Specific models are defined by their geometry (i.e. bond lengths and angles), Lennard-Jones parameters (σ and ϵ), and charges. The water models that I have used are part of the TIP family, first developed by Jorgensen [68], and are TIP3P [69] with modifications for the CHARMM force field [70, 71] and TIP4P-D [64]. The TI4P-D model uses the same geometry as the preceding TIP4P/2005 model [72], but has increased dispersion interactions (part of the Lennard-Jones interactions), aimed at sampling more extended conformations of IDPs. Another set of three-site models is the SPC family. The key difference between TIP and SPC is the geometry of the water molecule, which in TIP closely approximates experimental values (bond length $l = 0.9572$ Å and bond angle $\theta = 104.52°$), while the SPC water molecule mimics the tetrahedral shape of water molecules in ice ($l = 1$ Å and $\theta = 109.5°$) [67].

5.2.2 Force fields

The potentials described in section 5.2 together with the parameter set (e.g. force constants, equilibrium angles, and charges) constitutes a force field, which provides the foundation of a simulation. Although the dream is to have one force field that can describe all possible types of molecular systems, this is far from reality. Force field parameters are generally obtained from quantum chemical calculations and/or fitting with experimental data for a set of molecules, meaning that different force fields are aimed at different molecular systems. For proteins, the most widely used force fields families are Amber, CHARMM, GROMOS, and OPLS-AA. For a description of similarities and differences between these families, the reader is referred to ref. [73]. When discussing force fields, it is important to point out the relation to water models. Most force fields have been developed to work with a specific water model, and it has been shown that for IDPs even subtle changes in water model can influence the conformational ensemble sampled [74, 75]. Hence, it is important to use a correct combination of force field and water model.

While globular proteins and IDPs can appear indistinguishable at the most basic level; both being chains of amino acid residues connected by peptide bonds, standard force fields developed for globular proteins have been shown to work poorly for IDPs, by overestimating α-helical and β-strand structure [76–78] and producing overly compact conformations [79, 80]. Therefore, much effort has been put into improvements, resulting in numerous force fields [75, 78, 81–95]. For IDPs, there are mainly two types of improvements that have been relevant. The first is improvement of the propensity of sampling secondary structure, for example by adjustments of backbone dihedral parameters, such as in Amber ff03* and ff99SB* [82], and CHARMM22* [85]. Side-chain torsion potentials have also been improved, resulting in force fields like Amber ff99SB-ILDN [84]. Another approach with the same aim has been the introduction of energetic terms based on backbone dihedral cross-terms, so called grid-based energy correction maps (CMAP), first introduced in the CHARMM22/CMAP (CHARMM27) force field [81]. This force field was still shown to have bias towards α-helical structure, and therefore the CMAP potentials were refined against nuclear magnetic resonance (NMR) data, which together with updated sidechain dihedral parameters resulted in CHARMM36 [86]. Further refinement of CMAP potentials together with updates to Lennard-Jones parameters to correct arginine–glutamate/aspartate/C-terminus salt bridges, were introduced in CHARMM36m [75]. The second type of improvements has been aimed at overcoming collapse by balancing the protein–water and protein–protein interactions, for example by specifically targeting Lennard-Jones parameters between water and protein atoms as in Amber ff03ws [87], or by introducing a new water model [64]. A more profound description of force field development for IDPs can be found in the following reviews: [96–98].

As stated above, force fields generally perform best for systems that have been used in their optimisation. This also extends to the type of properties considered for validation. Hence, different force fields are better at reproducing some properties than others. Therefore, when selecting a force field, it is important to carefully consider the type of system and problem at hand, as well as perform tests and compare to experimental data.

Chapter 6

Simulation methods

Simulations act as a bridge between the microscopic and macroscopic world, and between theory and experiment. Through simulations we can obtain values of observables that can be measured in the lab, based on the interactions described in the model. In this way we can test a model by comparing with experiments, and test theoretical predictions on which the model is built. Given an accurate model, the simulations can also provide information not accessible by experiments.

In this work two different simulation methods have been employed: i) *Monte Carlo* (MC) to simulate the coarse-grained model and ii) *Molecular dynamics* (MD) to simulate the atomistic model. The main difference between MC and MD is that MC calculates ensemble averages based on random sampling, while MD is based on Newton's equations of motion, hence providing time averages. Recalling the first postulate of statistical mechanics stated in chapter 4, provided sufficiently long time and large ensembles, the result is the same.

6.1 Metropolis Monte Carlo simulations

As mentioned in chapter 4, the MC technique can be utilised to compute the ensemble average of an observable, given in Equation 4.8. In the simplest MC technique, often referred to as *random sampling*, this is done by evaluating the observable at a large number of random points in phase space and multiplying the result with the Boltzmann factor. Each point in phase space corresponds to a configuration. However, a lot of the generated configurations would only give a negligible contribution to the average, by having a really small Boltzmann factor. Such configurations are for example the ones in which particles are overlapping, since that results in a very high (or infinite) potential energy.

Metropolis et al. [99] presented a more efficient scheme for evaluating a ratio of integrals for obtaining the ensemble average. In this scheme the sampling is based on the Boltzmann factor, so that the sampling is focused more around configurations with a larger Boltzmann factor. This is a type of *importance sampling* and implies that the number of configurations needed for getting a good result is reduced, which makes the simulations faster. A Metropolis MC algorithm is outlined below [100]:

Metropolis Monte Carlo algorithm

i) Generate a starting configuration.

ii) Calculate the interaction energy within the system, U_{old}.

iii) Choose a particle at random and a type of trial move (see section 6.1.1).

iv) Generate a new configuration by performing the trial move on that particle.

v) Calculate the energy of the new configuration, U_{new}.

vi) Compare the energy of the old and the new configuration to determine if the new configuration is accepted. The probability of acceptance is given by:

$$p_{\text{acc}} = \begin{cases} 1 & \text{if } U_{\text{new}} \leq U_{\text{old}} \\ \exp[-\frac{1}{kT}(U_{\text{new}} - U_{\text{old}})] & \text{if } U_{\text{new}} > U_{\text{old}} \end{cases} .$$

vii) If the new configuration is rejected, restore the old one.

viii) Repeat from step ii.

To perform the MC simulations I have used the simulation package Molsim [101]. After an initial simulation allowing the system to equilibrate, the production run consisted of a single continuous run, divided into macrosteps, on which statistics have been calculated.

6.1.1 Trial moves

Trial moves are applied to generate new configurations of the system, to explore phase space. An advantage with Monte Carlo simulations is that unphysical moves can be used to speed up the exploration. In Paper I, four different types of moves, commonly applied to polymers and proteins modelled as bead-necklaces, were used. In Paper II I also implemented a cluster move, which is advantageous in self-associating systems.

Single particle translation: A single bead in the chain, or an ion, is moved to a new, ran-

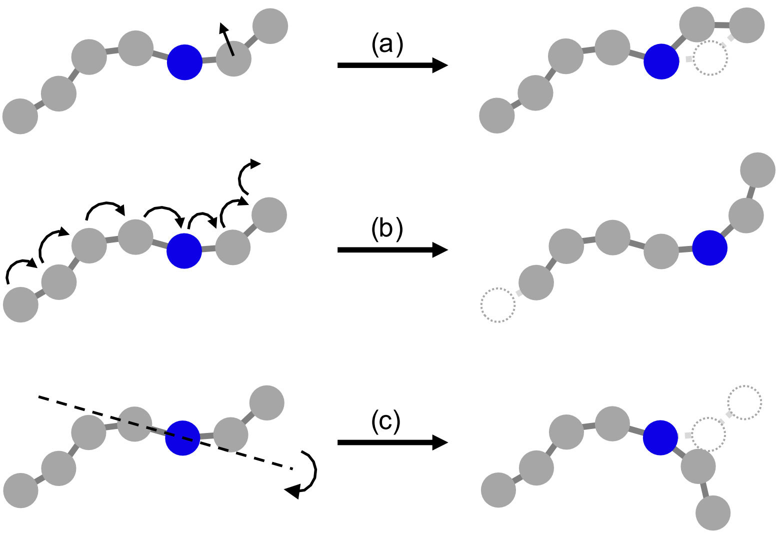


Figure 6.1: Illustration of three types of Monte Carlo moves: a) single particle displacement, b) slithering move, and c) pivot rotation.

domly chosen, position, see Figure 6.1a. The length of the translation is limited by an input parameter defined in the simulation.

Slithering move: In the slithering move, also known as reptation, one of the end beads is displaced to a random position within a bond length. The other beads are moved forward in the chain along the old configuration, as illustrated in Figure 6.1b.

Pivot rotation: One end of the chain is rotated around an axis defined by a randomly selected bond, see Figure 6.1c.

Chain translation: A whole chain is translated. This move does not change the conformation of the chain, only the position in relation to other chains and particles in the system.

Cluster move: A translation of a group of chains. The group includes the chain that the selected particle belongs to and all other chains whose center of mass is less than a predefined distance away. If the number of chains in the cluster changes during the displacement, the move is automatically rejected, as this violates detailed balance[1].

[1] Detailed balance implies that the probability of making a move and reversing it should be the same.

6.2 Molecular dynamics simulations

MD is another technique for computing equilibrium properties of classical many-body systems. In contrary to the MC technique, dynamical information can also be obtained due to the technique following Newton's equations of motion to move the particles. Newton's second law of motion states that for a particle i with constant mass, m_i, the force, $\mathbf{F}_i$ is proportional to the acceleration, $\mathbf{a}_i$, which can be expressed as the second derivative of the position $\mathbf{r}_i$ with respect to time, t:

$$\mathbf{F}_i = m_i \cdot \mathbf{a}_i = m_i \cdot \frac{\partial^2 \mathbf{r}_i}{\partial t^2}. \tag{6.1}$$

Hence, by knowing the forces, new positions and velocities of the particles can be generated by integrating Newton's second law of motion.

To run an MD simulation, starting velocities and positions, as well as the interaction potential are required as input. The forces are computed from the potential $U(\mathbf{r}^N)$, where $\mathbf{r}^N$ represents the complete set of atomic coordinates, according to

$$\mathbf{F}_i = -\frac{\partial U(\mathbf{r}^N)}{\partial \mathbf{r}_i}. \tag{6.2}$$

Since this is a many-body problem, we can only integrate the equations of motion numerically. Of course, the MD program relies on a good algorithm for doing this. I have used a version of the Verlet algorithm, called the *leap-frog algorithm*. In this algorithm, the velocities, $\mathbf{v}$, and the positions, $\mathbf{r}$, are updated at alternating times, as illustrated in Figure 6.2, using the following relations:

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) \tag{6.3}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \cdot \mathbf{v}(t + \frac{1}{2}\Delta t). \tag{6.4}$$

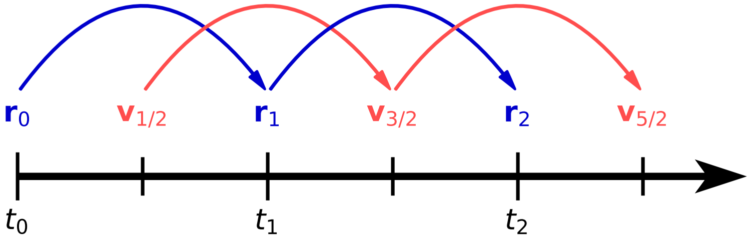


Figure 6.2: Schematic illustration of the leapfrog algorithm. It is called leapfrog due to the positions, **r**, and velocities, **v**, leaping over each other like frogs.

This algorithm is time reversible and area preserving, which contributes to its good energy-conserving properties. In addition, the algorithm allows for fairly long time steps, which is desirable since the number of time-consuming force evaluations then can be reduced [100, 102].

By repeatedly calculating the forces, velocities and positions, a trajectory showing how the positions and velocities changes with time is created. In this way, averages of observables can be obtained. A generic MD algorithm is summarized below [103]:

Molecular Dynamics algorithm

i) Initialize system: input the initial conditions (positions and velocities of all atoms in the system, and the potential interaction).

ii) Compute forces.

iii) Update configuration by numerically solving Newton's equations of motion.

iv) Write output.

v) Repeat from step ii.

For the MD simulations I have employed the GROMACS simulation package [58–62]. Each system has been simulated in several replicates, which have been initiated separately from the same structure, to obtain different starting velocities. Before final analysis, the individual replicates have been concatenated to one trajectory.

6.3 Technical details

In a simulation program, there are certain things that can be made to make the simulations more efficient or represent the system that we want. Here, some of those are described.

6.3.1 Periodic boundary conditions

Since this thesis investigates the behaviour of IDPs in solution, the simulations are supposed to represent bulk properties. Simulation systems can however not be as large as what is used in experiments, because that would entail an extremely large number of particles. For example, considering the most dilute samples in Paper II, even a small sample volume such as 0.1 mL contains about 10^{15} protein molecules, which is way too computationally demanding even for a coarse-grained simulation with implicit water. Unfortunately, the

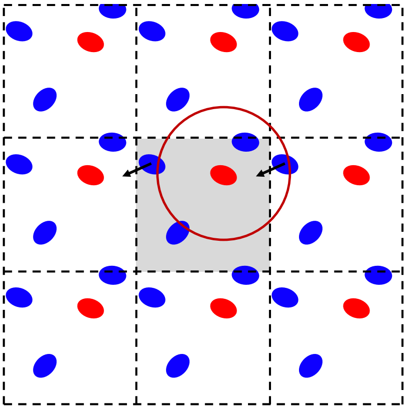


Figure 6.3: A schematic illustration of periodic boundary conditions in two dimensions, where the gray box is replicated in all directions. The arrows represents movement over a border. The red circle represents a spherical cut-off compliant with the minimum image convention for the particle marked in red.

relatively small system size employed in simulations causes a large part of the molecules in the system to be in contact with the walls of the box enclosing the system. Hence, to represent bulk behaviour, we employ periodic boundary conditions (PBC). This means that the simulation box is replicated in all directions to create an infinite lattice, as illustrated in Figure 6.3. In practice, this is achieved by letting a particle that leaves from one side of the box enter again from the opposite side. With this approach there are no walls in the system, hence it resembles the bulk. However, the periodicity of such a system can give rise to artefacts, especially if the simulation box is too small. Therefore, it is good practice to try different box sizes for the system. In the MD simulations, to ensure that the protein is not interacting with one of the periodic images, I have monitored the shortest distance between the protein and its closest periodic image. This distance should not fall below the cut-off applied to the non-bonded interactions. Cut-offs are further described in section 6.3.2.

In the coarse-grained simulations, a cubic box was employed, which is one of the simplest shapes that can be applied. However, in atomistic simulations using explicit solvent, a cubic box is not very efficient, due to the amount of solvent molecules needed to fill the corners of the cube. While a sphere is the most efficient volume, it cannot be combined with PBC. A shape that both has a smaller volume for the same image distance compared to a cube and is applicable for PBC is the rhombic dodecahedron, which has been used in the atomistic simulations.

6.3.2 Truncation

When dealing with an infinite system such as when using PBC, adding all the interactions in the system would lead to an infinite sum, due to the infinite number of particles. So for it to work practically, the interactions need to be truncated. Another reason for using truncation is that it increases the speed of the simulations, by reducing the number of calculations of non-bonded interactions. One approach is to use the minimum image convention, which restricts each molecule to interact only with the closest image of the other molecules. In practice, a spherical cut-off is often used, as illustrated in Figure 6.3. For a cubic box, the cut-off distance should not exceed half the box length, to comply with the minimum image convention. Truncating the interactions is often permissible dealing with short-ranged interactions, as the cut-off can be chosen sufficiently large, such that the interaction potential is zero beyond the cut-off. However, for long-ranged interactions, the contribution from the tail of the potential beyond the cut-off is usually non-negligible. Hence, to avoid errors, another approach is needed.

6.3.3 Long-range force handling

Due to the reasons described above, long-ranged electrostatic interactions are usually handled by the *particle-mesh Ewald* (PME) method [104], which is an improved version of Ewald summation. In Ewald summation the long-ranged interaction is separated into two parts: a short-ranged part treated as a direct sum, and a long-ranged part treated as a summation in reciprocal space. In this way, both parts converge rapidly. However, the computational cost scales as N^2, which makes it unsuitable for large systems. In PME, the reciprocal sum is approximated by a multidimensional piecewise interpolation. The approximate reciprocal energy and forces are expressed as convolutions and can therefore be evaluated using fast Fourier transforms, reducing the order of the algorithm to $N \cdot \ln N$, which makes it substantially faster than the original Ewald summation.

6.3.4 Neighbour lists

By employing cut-offs, the simulation program is sped up since the number of calculations of non-bonded interactions is reduced. However, iterating over all particles to calculate the distance between them, so that it can be determined which particles are within cut-off distance, still takes computational time. In liquids, it is usually the same particles that are in close vicinity over a few simulation steps, since it takes some simulation steps for the particles to move further away. By keeping lists over which particles are close, so-called *neighbour list*, we can avoid doing these calculations in every step. Due to having a ”buffer zone” outside the interaction cut-off when creating the neighbour lists, they can be updated less frequent. For a description of different ways to generate neighbour lists, the reader is referred to ref. [100].

6.3.5 Bond constraints

Another way to reduce the computational cost of MD simulations is by using a longer time step. The size of the time step is constrained by the time scale of the highest frequency motion in the system, which is usually bond vibrations of bonds involving hydrogen, limiting the time step to around 1 fs. Using a longer time step potentially makes the simulations unstable [105]. However, biomolecular simulations usually require simulation times in the order of µs–ms, which has a very high computational cost in terms of resources and/or physical time. By applying constraints on the bonds, such as by the LINCS algorithm [106], the length of the time step can be increased.

6.3.6 Controlling temperature and pressure

Direct use of MD simulations corresponds to the microcanonical (*NVE*) ensemble, since the Verlet-type integrators naturally conserves energy (assuming an appropriate time step). However, other ensembles can be a more convenient choice, for example the isothermal-isobaric (*NpT*) ensemble, having constant pressure and temperature, corresponding to the conditions of many laboratory experiments. The temperature and pressure can be controlled by applying temperature and pressure couplings. While there are several different options available, the *velocity-rescaling thermostat* [107] and the *Parrinello-Rahman barostat* [108] have been used for the MD simulations in this work.

The velocity rescaling thermostat is based on the Berendsen thermostat [109], in which the system is weakly coupled to an external heat bath, fixed at a desired temperature, T_0. The velocities of the particles in the system are rescaled in such a way that the rate of temperature change is proportional to the difference in temperature between the bath and the system:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}. \tag{6.5}$$

Here, τ is a time constant determining how strong the coupling is. A problem with the Berendsen thermostat is that it suppresses the fluctuations of the kinetic energy, meaning that it does not generate a proper canonical ensemble, hence the sampling is incorrect. In the velocity-rescaling thermostat this is corrected by an additional stochastic term that ensures a correct kinetic energy distribution [103]. When applying the Parrinello-Rahman barostat, additional terms involving the box vectors are included in the equations of motion, allowing the volume and shape to fluctuate.

Chapter 7

Simulation analyses

To characterise the simulated protein systems and obtain data that can be compared with experiments, I have performed different analyses, out of which the most important are described below.

7.1 Size and shape

The *radius of gyration*, R_g, is generally used as a measurement of size and is calculated as

$$R_\mathrm{g} = \sqrt{\frac{\sum_{i=1}^{n} m_i ||\mathbf{r}_i - \mathbf{r}_\mathrm{com}||^2}{\sum_{i=1}^{n} m_i}} \tag{7.1}$$

where m_i is the mass of element i, $\mathbf{r}_i$ the position of element i, $\mathbf{r}_\mathrm{com}$ is the center of mass, and n the total number of elements. In the atomistic simulations the elements are the atoms, while in the coarse-grained simulations they are the beads, with each bead having equal mass.

The *end-to-end distance*, R_ee, provides the distance between the N- and C- terminus and is given by

$$R_\mathrm{ee} = \sqrt{||\mathbf{r}_1 - \mathbf{r}_n||^2}, \tag{7.2}$$

where $\mathbf{r}_1$ and $\mathbf{r}_n$ is the position of the first and last element, respectively.

Defining the shape factor as

$$r_\mathrm{s} = \frac{R_\mathrm{ee}^2}{R_\mathrm{g}^2}, \tag{7.3}$$

we obtain a measurement of the shape of the IDP. For a Gaussian chain, r_s is approximately six, while in the rod-like limit it reaches twelve.

7.2 Scattering curves

For a direct comparison between experiments and simulations, scattering curves are measured by SAXS and corresponding curves are calculated in the simulations. The theory behind SAXS can be found in section 8.2. The scattering curves are calculated differently in the coarse-grained and the atomistic simulations, and will be presented separately.

7.2.1 Coarse-grained approach

Each particle (bead) is regarded as a point scatterer. For a system containing N identical scattering objects, the total structure factor is expressed as

$$S(\mathbf{q}) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right|^2 \right\rangle, \tag{7.4}$$

where $\mathbf{q}$ is the scattering vector. $S(\mathbf{q})$ can be further decomposed into partial structure factors given by

$$S_{jk}(\mathbf{q}) = \left\langle \frac{1}{(N_j N_k)^{1/2}} \left[\sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right] \left[\sum_{k=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_k) \right] \right\rangle, \tag{7.5}$$

where j and k are particle types. The total and partial structure factors are related through

$$S(\mathbf{q}) = \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} \frac{\left(N_j N_k\right)^{1/2}}{N} S_{jk}(\mathbf{q}). \tag{7.6}$$

For identical homogeneous spheres, the scattering intensity can be expressed as a product of the form factor and the structure factor, where the form factor corresponds to intra-particle interference and the structure factor to inter-particle interference. For a point scatterer, the form factor is constant, inferring that the scattering intensity is proportional to the structure factor. Consequently, the calculated structure factor for the point scatterers corresponds to the system's scattering intensity, only lacking a constant scaling factor. If the system is composed of a single protein chain, the calculated scattering profile comes only from intra-chain interference, hence, it is the protein form factor. For comparison with experiments an approximate effective particle form factor needs to be accounted for. This can be solved by dividing both the experimental and calculated scattering profile by their forward scattering, I_0.

7.2.2 Atomistic approach

There are several methods available for calculating solution scattering curves of macromolecules from atomic coordinates, of which the main differences regard the treatment of the solvent. The solvent is of importance because in a SAXS experiment, it is the excess electron density compared to pure solvent that is measured, meaning that the collected pattern corresponds to both the protein and the more dense layer of water molecules surrounding the protein, called the *hydration shell* (or hydration layer).

In this work I have used CRYSOL [110], in which the solvent is treated as a continuous electron density. The hydration shell is a 3 Å thick border layer with a constant excess electron density. The contrast of this hydration shell, i.e how much higher the water density is in this layer compared to the bulk, largely influences the calculated scattering curve. The effect of the contrast is especially evident in the Kratky plot, which provides information about the shape of the macromolecule. Unfortunately, choosing the optimal value of this contrast is not straightforward, as it has been shown to depend on both protein and force field [111]. A more robust way of obtaining the scattering curve is through explicit-solvent methods such as WAXSiS [112], which eliminate free parameters describing the hydration shell. However, it is associated with a higher computational cost.

7.3 Complex analyses

In Paper II, studying the self-association of statherin, several analyses are performed to characterise the result of the self-association, that is, the formed complexes. In these analyses, two chains are regarded as being part of the same complex if the center-to-center distance between a bead in each chain is less than a certain cut-off.

The complex size probability distribution is calculated according to

$$P_n = \frac{n\left\langle N_n^{\text{complex}}\right\rangle}{\sum\limits_n n\left\langle N_n^{\text{complex}}\right\rangle}, \tag{7.7}$$

where $\left\langle N_n^{\text{complex}}\right\rangle$ is the average number of complexes consisting of n chains [113]. Since the number of chains is constant in the simulations, the denominator is equal to the total number of chains in the system. Note that the distribution is weighted by the number of chains in each complex. The average association number is calculated from the complex size probability distribution, as

$$N_{\text{assoc}} = \sum_n nP_n. \tag{7.8}$$

To set the strength of the short-ranged hydrophobic interaction, in addition to comparing the average association number with experimental results, the number of contacts for each chain was monitored along the simulation. The purpose of that was to avoid a too large interaction, which would have prevented chains in complexes from separating. The geometric condition mentioned above was used to determine if two chains were in contact.

The shape of the complexes is determined from the the principal moments of the gyration tensor. For a perfect sphere, all three principal moments are equally large. The gyration tensor is calculated from the *x*, *y* and *z*-coordinates according to

$$S = \frac{1}{N}\begin{pmatrix} \sum_i^N X_i^2 & \sum_i^N X_iY_i & \sum_i^N X_iZ_i \\ \sum_i^N X_iY_i & \sum_i^N Y_i^2 & \sum_i^N Y_iZ_i \\ \sum_i^N X_iZ_i & \sum_i^N Y_iZ_i & \sum_i^N Z_i^2 \end{pmatrix}, \tag{7.9}$$

where $X_i = (x_i - x_{\text{com}})$ and similarly for Y and Z, and N is the number of beads in the complex. Through a transformation to a principal axis system such that

$$S = \text{diag}(R_1^2, R_2^2, R_3^2) \tag{7.10}$$

S is diagonalised and $R_1^2 \geq R_2^2 \geq R_3^2$ are the eigenvalues of S, also called the principal moments of the gyration tensor [114]. In the simulations the ensemble averages of the eigenvalues are calculated for each complex size separately. From the principal moments of the gyration tensor, the asphericity, α_s, is calculated according to

$$\alpha_s = \frac{\left(\langle R_1^2\rangle - \langle R_2^2\rangle\right)\left(\langle R_2^2\rangle - \langle R_3^2\rangle\right)\left(\langle R_3^2\rangle - \langle R_1^2\rangle\right)}{2\left(\langle R_1^2\rangle + \langle R_2^2\rangle + \langle R_3^2\rangle\right)^2}. \tag{7.11}$$

The asphericity ranges between 0 and 1, the values for a perfect sphere and a rod, respectively [115].

7.4 Secondary structure

In the coarse-grained model, no information regarding secondary structure of the IDP is available, since that requires finer details. However, from atomistic simulations, secondary structure can be determined. The program DSSP [116] calculates secondary structure based on hydrogen bonding patterns. Hydrogen bonds are defined through an electrostatic interaction energy between C=O and N–H groups, employing a generous cut-off. Secondary structure types that lack hydrogen bonding, such as bends, are determined based on geometric conditions. The secondary structure types defined in DSSP are α-helix, β-bridge, β-sheet, 3_{10}-helix, π-helix, hydrogen bonded turn, and bend. Residues not fulfilling the criteria for any of the aforementioned types are classified as having irregular structure. In IDPs, PPII structure is also common, which can be identified by DSSP-PPII [7, 117], an extension to the DSSP program. The DSSP-PPII program acts solely on what DSSP has classified as irregular, and uses a definition of PPII based on dihedral angles.

There are many available programs for secondary structure assignment, although DSSP is one of the most used. Another wide-spread program is STRIDE [118], which uses both hydrogen bonding patterns and dihedral angles. In the visualization tool VMD [119], secondary structure is assigned by STRIDE. Although DSSP and STRIDE often are in good agreement for structured proteins, especially in the assignment of α-helix and β-sheet, disagreement is somewhat larger among IDPs, where structural elements are usually shorter and more distorted. Differences are largest among turns, where DSPP and STRIDE use different definitions [120].

Experimentally, we have used circular dichroism (CD) spectroscopy to probe secondary structure. As will be discussed in section 8.3.2, it is challenging to obtain reliable quantitative measurements of secondary structure for IDPs from CD data. However, as an alternative, there are algorithms available that can calculate CD spectra from atomic coordinates [121, 122]. Such an algorithm can therefore be used to calculate the CD spectra from simulations, to compare with experimental data. However, recent studies have suggested that they are currently not reliable for IDPs [123, 124].

7.5 Salt bridges

In proteins, salt bridges can form between oppositely charged amino acid residues. In terms of intermolecular interactions, a salt bridge is a combination of an attractive charge–charge interaction and a hydrogen bond. Phosphorylated residues have the ability to form salt bridges with positively charged residues, and as Papers III–V show, this can greatly influence the conformational ensemble. We analyse salt bridges between phosphorylated and positively charged side groups based on formed hydrogen bonds, defined according to the Wernet-Nilsson criterion [125],

$$r_{\mathrm{DA}} < 3.3\ \text{Å} - 0.00044 \cdot \theta_{\mathrm{HDA}}^2, \tag{7.12}$$

where r_{DA} is the distance between donor and acceptor heavy atoms, and θ_{HDA} is the angle made by the hydrogen, donor, and acceptor atoms, given in degrees, with zero corresponding to a perfectly straight bond.

7.6 Principal component analysis

An important part of characterising IDPs is to get a view of the conformational ensemble. The complete energy landscape contain all information about a molecule and is described by $3N - 6$ internal coordinates, where N is the number of atoms of the system [126]. For most systems, this is a huge number of dimensions, making it impossible to handle. Additionally, the information content of a complete energy landscape is much larger than what we are interested in. Usually the goal is to find a few conformational classes, or arrive at a low-dimensional energy landscape that captures the relevant behaviour of the system in only a small set of coordinates. For this, *principal component analysis* (PCA) can be applied. It is a mathematical method for reducing the dimensionality of data while still retaining most of the variability, i.e. information content. PCA transforms the data from the original set of possibly correlated variables, into a new set of uncorrelated variables called principal components. The principal components are constructed as linear combinations of the original variables, in such a way that the first principal component accounts for as much of the variation of the data as possible. Each succeeding principal component account for as much of the remaining variation as possible, while still being orthogonal to the preceding components [127]. Hence, the information content is largest in the first few components, which makes it possible to scrap the remaining components and still retain a reasonable description of the system.

To construct low-dimensional energy landscapes of the IDPs in atomistic simulations, we follow the Campos and Baptista approach [126], where PCA is applied to the cartesian coordinates of the backbone atoms of the protein, obtained after translational and rotational least square fitting on a reference structure. Due to IDPs lacking an experimental reference structure, the central structure of the simulation, i.e. the conformation that differs least from all the sampled conformations, is used as reference. In mathematical terms, it is the conformation i among N sampled conformations that minimizes the dispersion measure

$$D_i = \left(\frac{1}{N-1} \sum_{j}^{N} \mathrm{RMSD}_{ij}^2 \right)^{1/2}, \tag{7.13}$$

where RMSD_{ij} is the root mean square deviation between backbone conformations i and j. After PCA, the probability density function, $P(\mathbf{r})$, in the representation space is estimated using a Gaussian kernel density estimator. The conditional free energy is then calculated according to

$$E(\mathbf{r}) = -RT \ln \frac{P(\mathbf{r})}{P_{\max}}, \tag{7.14}$$

where $P_{\max}$ is the maximum value of $P(\mathbf{r})$. This corresponds to assigning zero energy to the maximum of the probability density. The resulting energy landscape can be used to

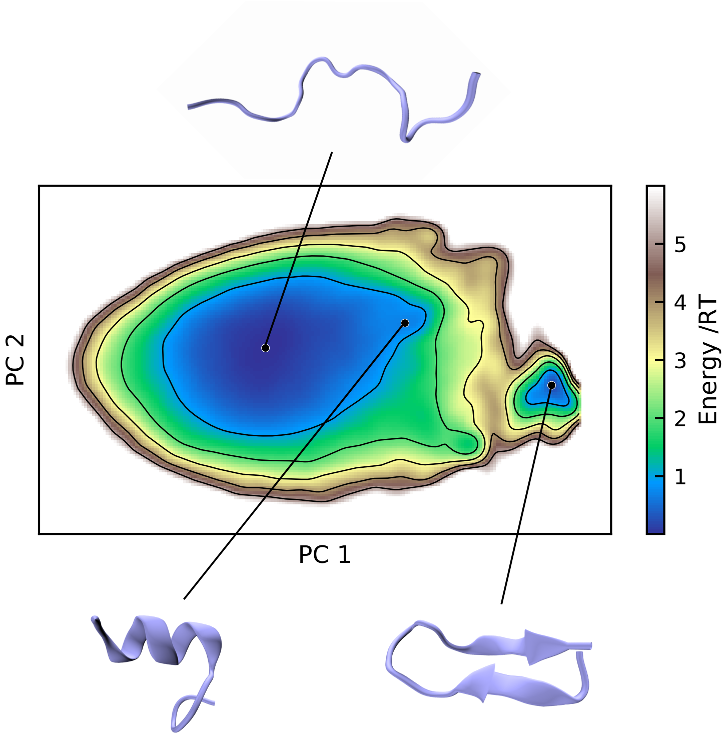


Figure 7.1: Schematic representation of an energy landscape constructed from the first two principal components that can be used to identify conformational classes.

compare different simulations and identifying conformational classes, as exemplified in Figure 7.1. However, for a complete picture of the conformational classes, more than the first two principal components are often required. This is due to that the major groups of conformations not necessarily are arranged in a non-overlapping way in this subspace, despite the first two principal components accounting for most of the variation.

7.7 Quality of sampling

In molecular simulations, there are two main factors causing errors: i) inaccurate models, and ii) insufficient sampling [128]. Hence, to be able to trust the simulation results and accredit discrepancies between simulations and experiments to model inaccuracies, we need to ensure proper sampling. It is important to keep in mind that it is much easier to rule out proper sampling than to prove it. In addition, without previous knowledge of phase space, there is no way to ensure that all important regions have been visited. Hence, focus needs to be on assuring good quality sampling in the regions visited. Here I will describe the methods used in this work, while a more profound guide can be found in for example these references [128, 129].

To check that basic equilibration has occurred, the time series of single observables can be observed, such as R_g and R_ee. For IDPs which exhibit a wide range of interchanging conformations, these observables usually show large fluctuations, however, systematic changes can often still be detected. The quality of sampling of single observables can be assessed by observing correlation and calculating error estimates. For a time-ordered series of values of an observable $f(t)$, the *auto-correlation function* at a time separation t' is given by

$$c_f(t') = \frac{\langle (f(t) - \langle f \rangle)(f(t+t') - \langle f \rangle) \rangle}{\sigma_f^2}, \tag{7.15}$$

where angular brackets denote the arithmetic mean, and σ_f^2 is the variance calculated as

$$\sigma_f^2 = \frac{1}{N-1} \sum_{i=1}^{N} (f_i - \langle f \rangle)^2, \tag{7.16}$$

where N is the number of values sampled. The auto-correlation function starts at one and decays towards zero as the correlation between values diminishes, i.e the simulation looses memory of earlier values. The time it takes for the simulation to loose memory is called the *correlation time*, and is more rigorously defined as

$$\tau = \int_0^\infty c_f(t')\mathrm{d}t'. \tag{7.17}$$

From the correlation time, it is possible to estimate the number of statistically independent values as the total simulated time divided by the correlation time, which can be used as a measurement of the quality of sampling of the observable. As a rule-of-thumb, the number of statistically independent values should be at least around 20 for the sampling of that observable to be considered reliable.

In *block averaging*, the trajectory is divided into M blocks of length n. For each block, the average of the observable, B_i, is calculated, yielding a total of M values. The block size n is gradually increased, and for each block size, the block-averaged standard error is calculated as

$$\mathrm{BSE}(n) = \frac{\sum_{i=1}^{M} (B_i - \langle B \rangle)^2}{M(M-1)}, \tag{7.18}$$

where $\langle B \rangle$ is the total average for the given block size. When the block length is substantially larger than the correlation time, i.e. the blocks are independent of each other, the BSE is a reliable estimator of the true standard error. For very small block sizes, when the consecutive blocks are highly correlated, BSE greatly underestimates the statistical error. Hence, BSE(n) increases with n until it reaches an asymptote to the true standard error. A converged BSE plot therefore signalizes that the error estimate for that observable has converged.

While the described methods above provides information about the sampling of single observables, it says little about the global sampling quality, i.e. how well the conformational space is sampled. Therefore, best practice is to always run several replicates with different initial conditions to compare.

Chapter 8

Experimental methods

In order to ensure that the simulation models describe the real world, we need to evaluate them against experimental data. Some of the most common techniques for experimental studies of IDPs are SAXS, single-molecule fluorescence resonance energy transfer (smFRET), and NMR, which all provide ensemble averaged data. This chapter focuses on the experimental techniques applied in this work, namely SAXS and CD spectroscopy. First however, I give a description of my protein purification process. In contrary to simulations were we are in complete control over what is included in the simulation box, real-world products purchased are never 100% pure. Therefore, the sample preparation and especially the protein purification is an important step in every experiment. In addition, the last section highlights some things to be aware of when using experimental data as validation.

8.1 Protein purification and determination of concentration

Statherin and the peptide fragments used in this work were purchased as lyophilised powders. The statherin powder contained trifluoroacetate, which lowered the pH, so that small addition of sodium hydroxide was necessary to dissolve the protein in buffer. To remove impurities and other buffer remains, the proteins and peptides were purified by two alternative methods. In the first, the protein solution was rinsed with buffer corresponding to at least 30 times the final sample volume, by centrifugation at a maximum speed of 358g at 8 °C in concentration cells with a 2 kDa cutoff. In the second method, dialysis was performed in room temperature and at 6 °C against a buffer of at least 400 times the sample volume, using 0.5–1 kDa membranes and exchanging the buffer 4 times during 48 h.

In both SAXS and CD experiments, the recorded signal depends on the protein concentration. Hence, for processing and interpreting the data it is important to know the concentration. I have determined the concentration by absorption measurements using a Nanodrop 2000 spectrometer. For statherin, measurements were performed at 280 nm using an extinction coefficient of 8740 $M^{-1}cm^{-1}$. Since the 15 residue long N-terminal fragment of statherin lacks residues with aromatic rings, measurements were instead performed at 214 nm, using an extinction coefficient of 24000 $M^{-1}cm^{-1}$, calculated based on contributions of the peptide bond and the individual amino acids present, according to Kuipers and Gruppen [130]. In Paper III, due to limitations posed by available equipment, the concentration of the statherin fragment samples for SAXS were determined at 257 nm, where phenylalanine absorbs. The extinction coefficient used was 390 $M^{-1}cm^{-1}$, based on the value reported by Mihalyi [131]. However, here the absorption was rather low, so this approach was associated with a larger uncertainty.

8.2 Small-angle X-ray scattering

SAXS is a low-resolution technique commonly used to probe the average size, shape, and structure of particles in the nanometer length scale, typically between 1 and 100 nm. It can be applied to samples in different states such as liquid and solid, but here we focus on solution scattering of biological macromolecules.

8.2.1 Basic principle

In a SAXS experiment, a narrow beam of X-rays is sent through a sample. The X-rays interact with the electrons in the atoms, which causes the atoms to emit spherical scattered waves. The scattered waves interfere, which gives rise to an interference pattern at the detector, from which structural information can be extracted. A schematic set-up of the main parts of a SAXS instrument is found in Figure 8.1.

Scattering can occur with or without the loss of energy, however, it is the elastic scattering,

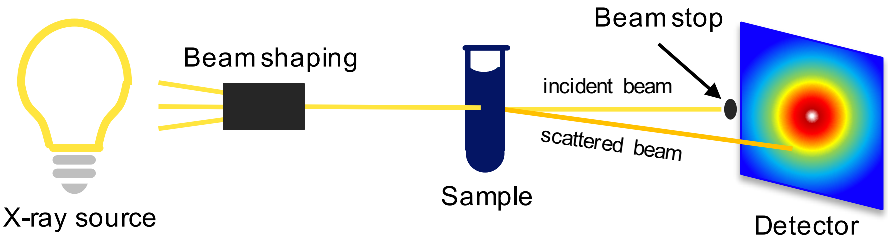


Figure 8.1: A schematic representation of the main components in a SAXS instrument. The beam stop hinders the incident beam from reaching the detector and overshadowing the sample scattering.

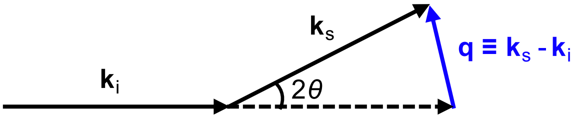


Figure 8.2: A schematic representation of the scattering vector **q**, defined by the incident wave vector $\mathbf{k}_i$ and the scattered wave vector, $\mathbf{k}_s$.

that occurs without energy loss, that is of importance for SAXS. Both the incident beam and the scattered beam can be considered as planar waves defined by a wave vector, $\mathbf{k}_i$ and $\mathbf{k}_s$, respectively. The momentum transfer, usually referred to as the scattering vector, **q**, is defined as the difference between the incident and scattered wave vectors, as illustrated in Figure 8.2. The magnitude of the incident wave vector is $||\mathbf{k}_i|| = 2\pi/\lambda$, where λ is the wavelength of the incident beam. Since there is no loss of energy in elastic scattering, $||\mathbf{k}_s|| = ||\mathbf{k}_i||$, hence, the magnitude of **q** can be expressed as

$$q = \frac{4\pi}{\lambda}\sin(\theta), \tag{8.1}$$

where 2θ is the angle between the incident and scattered wave vector [132].

Since the X-rays are scattered due to interactions with electrons, the more electrons a sample contains, the stronger the scattering signal is. The difference in electron density throughout the sample is therefore responsible for creating the contrast. Biological macromolecules contain mostly light elements such as hydrogen and carbon, thus the difference in electron density compared to the aqueous solution is small. Hence, the resulting signal is especially weak [132]. Therefore, for biological samples, it can be advantageous to use X-rays produced from a synchrotron, a type of large circular accelerator, instead of a lab source. The synchrotron produces X-rays with much higher brilliance, which means that the exposure time needed for detecting a useful signal is much shorter, often a few seconds compared to hours. However, the risk of radiation damage to the sample is much higher. Therefore, several frames are recorded of each sample, to compare for radiation damage and collect statistics. Also, I have used Tris buffer, which acts as a radical scavenger and therefore reduces radiation damage, in contrary to phosphate buffer which can promote it [133].

8.2.2 The scattering intensity

The detector records the scattering intensity at positions in two dimensions, however, since thermal motion causes the orientation of the particles to be random in respect to the incident beam, the scattering signal is a spherical average and can therefore be reduced to one dimension. The scattering intensity is usually presented as a function of q, to be independent of the wavelength. When performing a SAXS experiment, the scattering of the full sample is recorded. To obtain the scattering curve of only the solute of interest, in my case the protein, we need to subtract the background. Therefore, the scattering of a matching buffer is also measured. A poorly matched buffer will greatly affect the data, so to ensure a good match, I dialysed all stock solutions overnight. The resulting dialysis buffers were used for background measurements and to dilute the samples into a concentration series.

The scattering intensity contains information on both the single particle (intraparticle interference) and relation between different particles (interparticle interference). Assuming the system consists of identical homogeneous spheres, the scattering intensity can be expressed as

$$I(q) = P(q) \cdot S(q), \tag{8.2}$$

where $P(q)$ is the form factor and $S(q)$ is the structure factor. From the form factor the size and shape of the individual particle can be determined. The structure factor contains information on the distance between particles, which can show if the particles are repelling or attracting each other. Attraction will increase the scattering curve at low q and repulsion will decrease it. In dilute and weakly interacting systems no structure is formed in the solution, meaning that the structure factor is a constant. Hence, at such conditions the form factor can be determined. Different form factors are illustrated in Figure 8.3a.

Note that IDPs adopt many different conformations, so the measured SAXS pattern corresponds to an average over all these conformations. Likewise, when dealing with polydisperse samples containing particles of different sizes, the resulting SAXS curve is an average over the different sizes present.

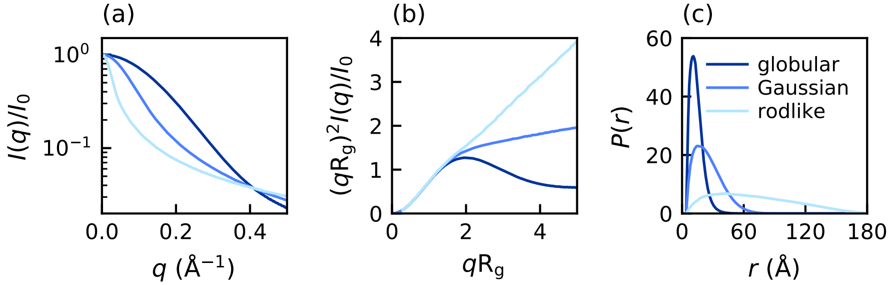


Figure 8.3: Illustration of the differences between a more globular, flexible (Gaussian chain-like) and rodlike protein. a) Form factor, b) dimensionless Kratky plot, and c) pair distance distribution function.

8.2.3 Data analysis

For proteins some standard analyses which do not require any modelling are usually performed. Besides providing information regarding particle shape and size, they also serve as a check of data quality.

The Guinier approximation

The Guinier approximation [134] provides a relation between the scattering curve at low q and the object size given by R_g, according to

$$\ln I(q) = \ln I_0 - (R_\mathrm{g} q)^2/3, \tag{8.3}$$

where I_0 is the forward scattering (the scattering signal extrapolated to $q = 0$). Usually $\ln I(q)$ is linear with respect to q^2 at small q, normally in the region $qR_\mathrm{g} < 1.3$ for well-folded proteins. For IDPs, this region can be reduced to $qR_\mathrm{g} < 0.8$ [135]. Using a too large q-range tends to underestimate the R_g. If the Guinier plot shows an upswing at low q this indicates considerable aggregation in the sample, while a downswing corresponds to intermolecular repulsion. In both cases the data quality is compromised and detailed analysis should be avoided.

The forward scattering is related to the molecular weight by

$$M_\mathrm{w} = \frac{I_0 \cdot N_\mathrm{A}}{c([\rho_\mathrm{p} - \rho_\mathrm{s}]\nu_\mathrm{p})} \tag{8.4}$$

where I_0 is given in absolute units (cm^{-1}) and c is the protein concentration. The electron density of the protein, ρ_p, the electron density of the solvent, ρ_s, and the partial specific volume of the protein, ν_p, can all be calculated theoretically. The forward scattering is measured in arbitrary units that differs between detectors, but can be transformed to absolute units, for example by measuring the scattering of water. Normally a difference less than 10% between the measured and the theoretical weight is regarded as good [54, 136]. For self-associating proteins such as statherin, the average association number can be calculated from the measured molecular weight. Note however that for a polydisperse sample, this average is not the number average. The scattering from a sphere can be expressed analytically, from which it can be shown that in the $q \to 0$ limit, $I \propto R^6$, where R is the sphere radius [132]. Hence, large particles contribute more to the average than small particles. This is also the reason why SAXS is so sensitive to aggregates in the sample. To remove possible large aggregates from the samples, I centrifuged all protein stock solutions at approximately 18000g for at least 2 hours, after which the bottom 1/3 of the samples were discarded.

Kratky plot

To assess the flexibility of a protein and differentiate between globular and disordered proteins the Kratky plot is useful. A dimensionless Kratky plot allows for comparison between proteins of different sizes, and is constructed as $(qR_\mathrm{g})^2 I(q)/I_0$ vs qR_g [137]. Figure 8.3b illustrates the different behaviour of a more globular, Gaussian chain-like and rodlike protein. An intrinsically disordered protein usually exhibits a plateau as the Gaussian chain, while the actual slope depends on for example the amount of partial structure.

Pair distance distribution function

The pair distance distribution function, $P(r)$, provides information on shape, since it shows the distribution of pair distances within the protein. It is expressed in real space, compared to the scattering pattern that contains information in inverse space. $I(q)$ and $P(r)$ are related by a Fourier transform, according to [132]

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(q) qr \sin(qr) \mathrm{d}q. \tag{8.5}$$

Since $I(q)$ is not known over the full interval $0 \leq q \leq \infty$, $P(r)$ can not be obtained directly, hence an indirect Fourier transformation method [138, 139] is often used. By definition, $P(r)$ is equal to zero at $r = 0$ and $r = D_\mathrm{max}$, the maximum distance within the protein. Since proteins do not have hard surfaces, the distribution is expected to approach zero smoothly. Problems of reaching zero or small peaks at larger r values are indicative of aggregation in the sample [140].

The $P(r)$ provides easy differentiation between globular and unfolded proteins, such as IDPs, as illustrated by Figure 8.3c. For a globular protein, the $P(r)$ is a symmetric bell-shaped curve, while for an unfolded protein the $P(r)$ shows an extended tail. If a protein has multiple domains it can be detected in the $P(r)$ as two different peaks.

R_g and I_0 can also be calculated from $P(r)$, by using the equations below [135]

$$R_\mathrm{g}^2 = \frac{\int_0^{D_\mathrm{max}} r^2 P(r) \mathrm{d}r}{2 \int_0^{D_\mathrm{max}} P(r) \mathrm{d}r} \tag{8.6}$$

$$I_0 = 4\pi \int_0^{D_\mathrm{max}} P(r) \mathrm{d}r. \tag{8.7}$$

Since the Guinier method only uses a small region of the scattering curve, while $P(r)$ is based on more or less the whole curve, the Guinier method is more susceptible to experimental noise, giving rise to larger uncertainties. Hence, the $P(r)$ method can be more reliable.

However, the Guinier method normally has better reproducibility between users, as it is an easier method to apply. Ideally, the R_g determined from both methods should be in agreement. Note however, that R_g determined from SAXS is not directly comparable to the R_g calculated in simulations using equation 7.1, due to the scattering pattern including contributions from the hydration shell surrounding the protein [111, 141].

8.2.4 Size exclusion chromatography-coupled SAXS

A size exclusion chromatography (SEC) column is used for separating a sample according to size. A SEC column usually contains porous beads that allow small molecules to travel into the bead pores, while large objects only moves in between the beads. Hence, smaller objects travel a longer route and will be eluted later than large objects. A SEC column can therefore be used in-line with SAXS to separate the sample according to size and measure SAXS directly as it is eluted. For polydisperse samples it is therefore possible be to obtain SAXS curves for the different sized objects individually and hence obtain a size distribution. SEC-SAXS is also useful in obtaining the form factor for samples prone to aggregate, since the aggregates and the monomeric protein are eluted at different times.

8.3 Circular dichroism spectroscopy

CD spectroscopy is a highly sensitive but low-resolution technique based on the adsorption of polarised light and provides information on the secondary structure content in proteins.

8.3.1 Basic principle

Light is a type of electromagnetic radiation, which comprises an electric field and a magnetic field. These fields oscillate in perpendicular planes, that also are perpendicular to the direction of propagation. Normally light is unpolarised, which means that it oscillates in all possible directions. In linearly polarised light, the oscillations are restricted to only one direction, as illustrated in Figure 8.4a. In circularly polarised light, the electric vector rotates around the direction of propagation, undergoing a full revolution per wavelength. Clockwise rotation corresponds to right circularly polarised light, and counterclockwise to left circularly polarised light [142].

Linearly polarised light can be viewed as made up by two components of circularly polarised light of equal magnitude and phase, rotating in opposite directions (left and right), as illustrated in Figure 8.4b. If the two components are of different amplitudes, the light will be elliptically polarised, as the electric vector instead will trace an ellipse, see Figure 8.4c.

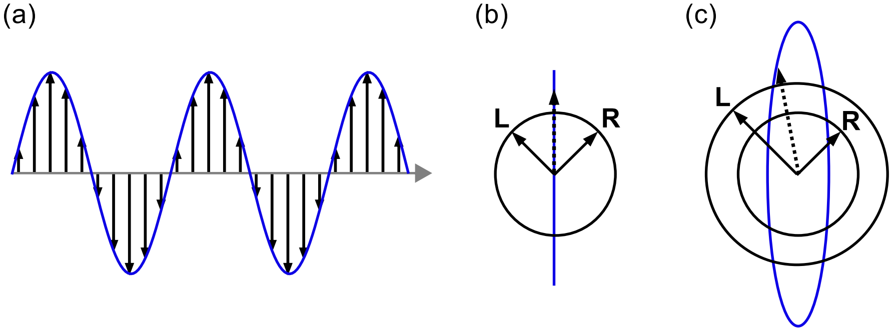


Figure 8.4: a) An illustration of linearly polarised light. The grey arrow corresponds to the direction of propagation and the black arrows represent the electric vector at different points along the propagation. b) Linearly polarized light made up by two components of circularly polarized light L and R rotating in opposite directions. The dashed arrow represents the electric vector and corresponding to the sum of the two components, which is always oriented along the blue line. (c) Different amplitude of the two components causes the electric vector (dashed arrow) to trace an ellipse, outlined in blue.

This is what happens during a CD spectroscopy experiment, as an optically active sample absorbs the left and right circularly polarised light to different extents [143].

An optically active sample contains chromophores, i.e. light-absorbing groups, that are chiral, covalently linked to a chiral centre, or situated in a chiral environment due to the three-dimensional structure of the molecule. In a protein, the chromophores of largest interest are the peptide bond, aromatic amino acid side chains and the disulphide bond. The far UV-region (approximately 170-250 nm) is dominated by peptide bond absorption, and it is in this region different secondary structure give rise to characteristic patterns, see Figure 8.5 [142].

8.3.2 Data analysis

A CD experiment monitors the difference in absorption of left and right circularly polarised light for different wavelengths. To ensure a good signal from the protein, the absorbance of the buffer should be low. Chloride ions strongly absorbs light at wavelengths in the lower end of the UV region of interest [143], and therefore I used sodium fluoride instead of sodium chloride in the CD samples. Also Tris absorbs in this region, so phosphate buffer was used instead. Aggregates and dust particles can create artefacts in the data [143], so all samples were filtered through a 0.22-µm hydrophilic filter before measurement.

Due to historic reasons the spectrum is usually presented in terms of ellipticity, with the unit degrees, and not as a difference in absorbance (ΔA). The ellipticity, θ, is calculated from the major and minor axes of the resulting ellipse and is related to the absorbance by $\theta = 32.98\Delta A$. The magnitude of the CD signal depends on the sample concentration

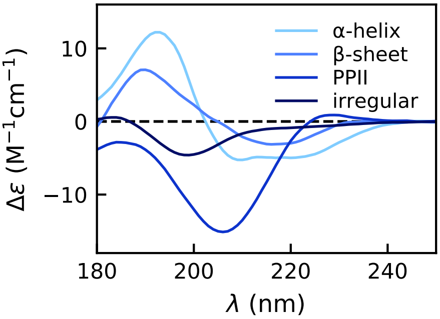


Figure 8.5: CD spectra of proteins with different secondary structure. The spectra are obtained from the Protein Circular Dichroism Data Bank [144] with the following spectrum id: CD0000117000 (α-helix) [145], CD0000118000 (anti-parallel β-sheet) [145], CD0004553000 (PPII) [146], and CD0006124000 (irregular) [147].

and the path length, so to be able to compare different measurements, the signal needs to be normalised. A common approach is to express the signal as the mean residue ellipticity (unit: deg·cm^2·dmol^{-1}), calculated as

$$[\theta]_{\mathrm{MRW}} = \frac{\theta \cdot \mathrm{MRW}}{10 \cdot d \cdot c}, \tag{8.8}$$

where θ is the observed ellipticity (in mdeg), d the path length of the cell (in cm), and c the protein concentration (in mg/mL). The mean residue weight, MRW, is the molecular weight (in Da) divided by the number of peptide bonds [143]. Data in absorption units is often expressed as the molar differential extinction coefficient, $\Delta\varepsilon$ (unit: M^{-1}cm^{-1}), calculated as

$$\Delta\varepsilon = \frac{\Delta A}{C \cdot d}, \tag{8.9}$$

where C is the molar concentration (in M).

By observing the shape of the CD spectrum, it is usually possible to discern the dominating type of secondary structure. Monitoring the shape is a straightforward method for detecting conformational and structural changes upon changes in environment, such as salt concentration or temperature. To obtain a quantified measurement of the secondary structure composition from the CD spectrum, there are several different methods available. They are all based on the approximation that a given protein CD spectrum can be expressed as a linear combination of spectra of different secondary structure components [148]. Hence, a good reference data set is vital to the results. A big reference set is often advantageous to account for some of the structural variability within a secondary structure type. Still, results can vary with both method used and applied reference set. Since irregular structure, sometimes referred to as random coil, is not a defined secondary structure, rather the lack of other structural elements, its variability is especially large. Hence, structural assessment of IDPs from CD data is particularly challenging. Furthermore, most methods are optimized for globular proteins, meaning the result for short peptides and IDPs can be questionable. It is therefore advantageous to compare the result of different methods and/or basis sets before drawing conclusions, or only use CD spectroscopy as an indicative tool of changes in secondary structure.

8.4 Using experimental data to evaluate simulation models

By using experimental data for investigating whether the simulation models are correct, we assume that the experimental data is representative of the real world. However, even when disregarding errors that can occur in the execution of experiments, as we have seen above, approximations and assumptions are often used in the processing of data. This of course affects the final data and is another possible source of discrepancy between simulations and experiments. It is therefore preferable if the observables measured in experiments can be calculated directly in simulations.

Something else to consider is that the methods described above are rather low in resolution and measure ensemble averages. This implies that it is easier to prove a model incorrect than correct, since for example a given SAXS curve can agree equally well with different ensembles of structures. Hence, best practice is to always use several experimental methods to compare with. Just as SAXS, smFRET provides information on the overall chain dimensions, by probing long-range distances within IDPs. Connecting fluorophores to the N- and C-terminus, R_{ee} can be determined by assuming a shape of the distance distribution based on polymer theory [149]. However, the necessary fluorophores have actually been shown to influence the conformational properties of the IDP, which needs to be corrected for [150]. NMR data in the form of chemical shifts and scalar couplings contain information about local-level phenomena such as secondary structure content, and have also been applied for force field validation [65, 77, 92, 151]. In fact, regarding atomistic simulations, it has been shown that overall chain dimensions and secondary structure content is largely independent of each other, such that experimental data of both types need to be used in proper validation of force fields [152].

Lastly, when comparing results of a simulation model to experimental data, we should be aware of the intended purpose of the model. Quantitative agreement with experimental data is not always required for a model to be useful. In fact, qualitative agreement through trends can be enough, depending on the research question asked.

Chapter 9

The research

This chapter summarises and discusses the papers compiling this thesis. Overall, the research has been focused on investigating models and force fields and explore the conformational ensembles of IDPs. The first two papers explored the coarse-grained "one bead per residue"-model. Paper I investigated the generality of the model in dilute conditions, while Paper II applied the model to the self-association of statherin. In Paper III–V focus was shifted to the role of phosphorylation, which required an atomistic approach to capture changes in secondary structure. Paper III studied the 15 residue long N-terminal fragment of statherin using two different force fields. The force fields were further evaluated in Paper IV for an additional four peptides, and in Paper V the most appropriate force field was used to investigate the conformational effects induced by phosphorylation.

9.1 The generality of the coarse-grained model at dilute conditions

To test the generality of the coarse-grained model, in Paper I MC simulations of a single chain with explicit counterions and implicit salt and water, were performed for the ten different intrinsically disordered proteins or regions summarized in Table 9.1. According to the Das-Pappu plot in Figure 9.1a, this selection of IDPs represent all four conformational classes of IDPs. Hence, although the number of IDPs studied is fairly small, they still provide a good representation.

The R_g determined from simulations were compared to the R_g reported from SAXS measurements at 150 mM. As Figure 9.1b shows, the simulated values were overall in rather good agreement with the experimental values, suggesting that the model can be applied to a range of different IDPs. However, for some sequences the simulated value was distinctly smaller than the experimental value, considering the reported uncertainty, namely

Table 9.1: Length, number of phosphorylated residues (N_{phos}), fraction of charged residues (FCR), net charge per residue (NCPR), proline content (Pro), and hydrophobic content (H-phob) of the IDPs studied in Paper 1. The name of the phosphorylated IDPs are printed in red, while yellow represents proline-rich IDPs.

IDP	Length	N_{phos}	FCR	NCPR	Pro (%)	H-phob (%)
histatin 5_{4-15}	12	0	0.42	+0.42	0	17
histatin 5	24	0	0.38	+0.21	0	8
statherin	43	2	0.28	-0.09	16	16
IB5	73	0	0.11	+0.08	40	7
ash1	83	0	0.20	+0.18	15	14
pash1	83	10	0.45	-0.06	14	14
sic1	92	0	0.12	+0.12	16	22
psic1	92	6	0.25	-0.01	16	20
II-1ng	141	0	0.19	+0.11	36	1
RNase E	248	0	0.39	+0.05	6	22

for pAsh1, pSic1, II-1ng, and RNase E. For RNase E it is plausible that the discrepancy was caused by a slight degree of self-association affecting the SAXS data. II-1ng is rich in prolines, which is known to increase stiffness. This effect has not been accounted for in the model, hence a smaller simulated value could be expected. The discrepancies for II-1ng and RNase E were however relatively small, compared to the discrepancies for pAsh1 and pSic1, which are most probably due to their high number of phosphorylated residues, which will be discussed later on.

Further-on, the experimental R_g could be fitted to a power law expression typical for polymers:

$$R_g = \rho_0 N^{\nu}, \qquad (9.1)$$

where ρ_0 is a prefactor, N is the number of monomers (i.e amino acid residues), and ν is the Flory exponent, determined to 0.59, which agrees with the value for a self-avoiding random walk (SARW), which is approximately 0.6. This indicates that this selection of IDPs can be approximated as SARWs under the experimental conditions used, namely high ionic strength (150 mM). Therefore, it suggests that the intramolecular interactions are dominated by electrostatic interactions, which are highly screened at 150 mM.

Using a model system without charges, resembling the SARW, it was shown that the range of R_g values sampled increased with chain length, implying a relation between the conformational entropy and chain length. For all chain lengths, the probability distribution of the shape factor was a broad bell-shaped curve ranging between zero and twelve (the rod-like limit) with a maximum value of 15% at six, the value for an ideal chain. This shows that IDPs indeed adopt a wide range of different conformations, so that the conformational ensemble description is necessary.

Since IDPs are generally rather sensitive to environmental changes due to their rather flat conformational landscapes, the effect of ionic strength is of interest. Indeed the number of

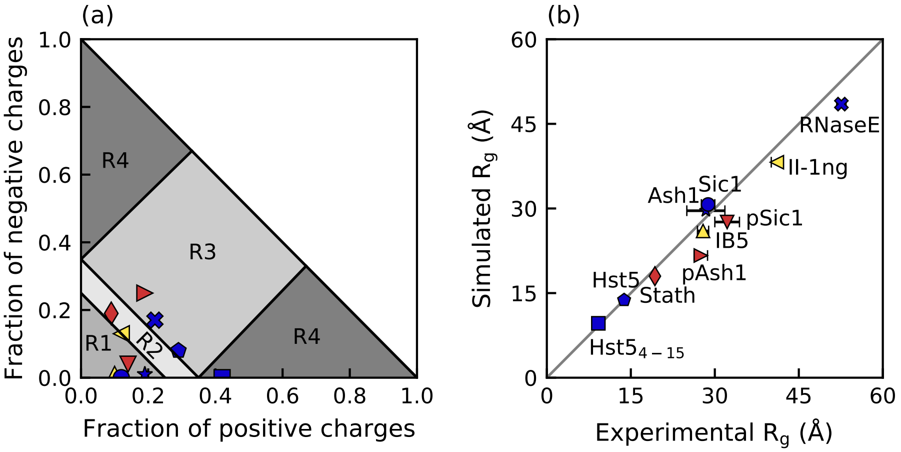


Figure 9.1: a) Classification of the IDPs included in Paper I according to the Das-Pappu plot. The regions are globules (R1), globules and coils (R2), coils/hairpins (R3), and coils/semiflexible rods (R4). Radii of gyration obtained from simulations *versus* the radii of gyration determined from SAXS experiments. In both panels proline-rich IDPs are shown in yellow, phosphorylated in red, and the rest in blue.

charged residues and their distribution throughout the sequence controlled the response to changes in ionic strength. For example, RNase E expanded upon increased ionic strength, in agreement with its classification as a strong polyampholyte, while Ash1 showed polyelectrolytic behaviour, i.e. a contraction. Although it was concluded that the IDPs could be approximated as SARWs at an ionic strength of 150 mM, Figure 9.2a confirms that this is an approximation. For Ash1, full agreement with the distribution of a SARW was reached first at 1000 mM, although the largest change occurred between 10 and 150 mM. In fact, the ionic strength was shown to have a considerable effect on the form factor. The form factor from simulations at both 150 mM and SARW conditions were in agreement with the experimental form factor collected at 150 mM NaCl, see Figure 9.2b,c. The form factor at 10 mM deviated, which implies that using the form factor collected at 150 mM salt to obtain the structure factor at 10 mM salt is indeed an approximation. However, depending on the system this approximation can be valid or contribute to errors.

To summarise, it appears that many IDPs can be described by this coarse-grained model including only steric contributions, electrostatic interactions and an approximate van der Waals interaction. The model is able to provide a basic understanding of the importance of chain length and charge distribution, and predict the outcome of changes in ionic strength. Of course, the model has its limitations. As pointed out above, the R_g of IB5 was slightly underestimated, and the stiffness shown by the Kratky plot as well. Including an angular potential made it possible to accurately represent the shape in accordance with the Kratky plot, however, this instead caused an overestimation of the R_g. To obtain a better repres-

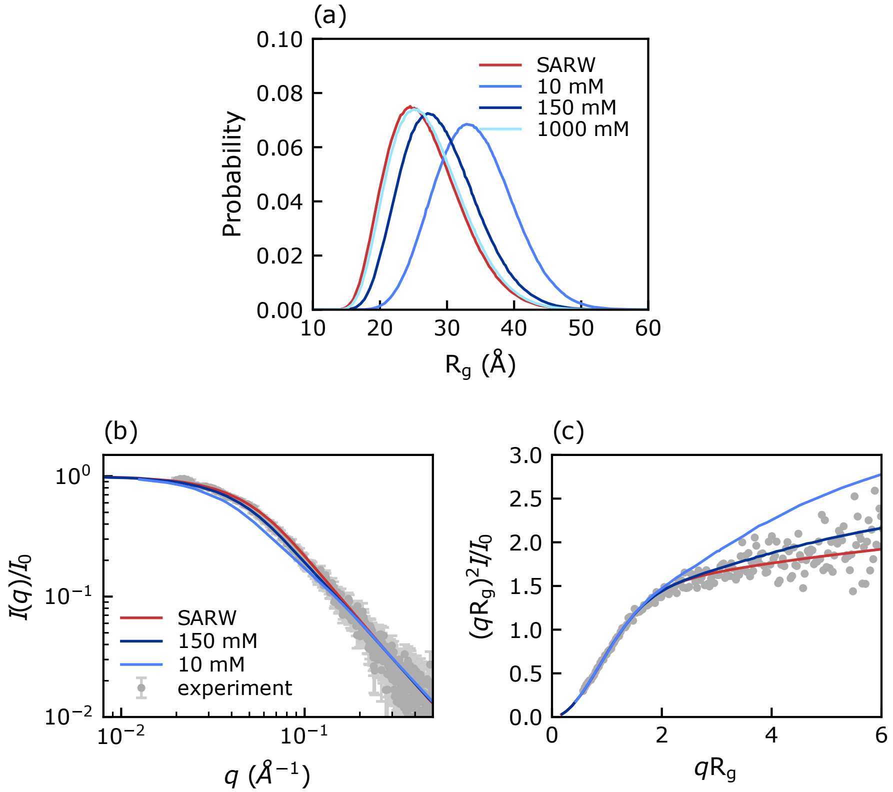


Figure 9.2: a) Probability distribution of the radius of gyration for Ash1. b) Form factor and c) dimensionless Kratky plot of Ash1 at 10 and 150 mM salt, and modelled as a SARW, compared to the experimental form factor collected at 150 mM NaCl, obtained from [153].

entation of both size and shape, a different approach, for example including local stiffness, would be necessary. The phosphorylated IDPs were also shown to be a challenge for the model. Statherin, the shortest and least phosphorylated of the three, showed a matching scattering curve and decent agreement of R_g, but for pSic1 and pAsh1 the model produced more collapsed ensembles than the experimental references. Interestingly, the agreement was much better using a charge of only $-1e$ on the phosphorylated residues. What appears as an overestimation of charges in the model may instead be caused by experimental deficiencies and/or errors and approximations within the model. For example, there can be a natural variation of the number of phosphorylated residues in the experimental sample, as well as traces of multivalent ions binding to some phosphorylated residues, meaning that the simulated and experimental sample might not be the same. Since the model has been parameterised by comparing with the form factor of histatin 5, the fact that the calculated R_g from simulations does not take into account a hydration shell, is not expected to cause discrepancy as long as the hydration shell is rather similar to that of histatin 5.

However, for Ash1/pAsh1 it was recently shown that the SAXS-derived R_g includes a larger hydration shell for the phosphorylated species, which makes it appear larger and therefore partly masks conformational changes induced by phosphorylation [141]. In addition, this model uses fixed charges, and it is possible that $-2e$ is an overestimation of the negative charge, considering the pK_a being approximately six [154] and possible influence from the local environment. As Section 9.3 will show, phosphorylation contributes with more than only charge–charge interactions, and these other factors can influence the conformational ensemble, such that a more detailed description than what this model provides might be necessary for an accurate description of phosphorylated IDPs.

9.2 Self-association of statherin

While Paper I showed that a coarse-grained model can be useful for exploring the conformational ensemble of IDPs at dilute conditions, one of the greatest benefits of a coarse-grained approach is that it enables studies of larger and more complex systems, where the computational load of an atomistic model is too large to be feasible. Hence, in Paper II the aim was to apply the model for understanding the balance between interactions in a self-associating IDP system. The saliva protein statherin was used as a model system, due to its amphiphilic character and relatively short chain length. Using SAXS, it was shown that statherin forms complexes upon increased protein concentration, see Figure 9.3a. The self-association ceased with the addition of 8 M urea, and diminished by increased temperature or lowered ionic strength. Changes in the Kratky plot (Figure 9.3b) and $P(r)$ showed that the formed complexes were more globular than the monomeric protein.

Although the exact mechanism of how urea affects proteins and self-associating systems has

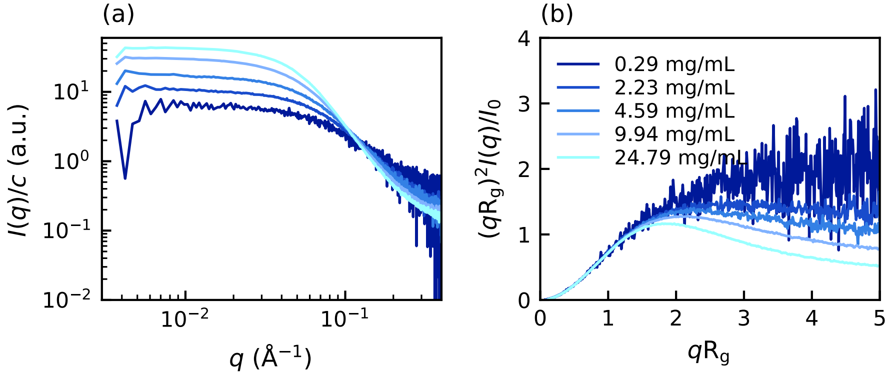


Figure 9.3: a) Scattering intensities and b) dimensionless Kratky plot of increasing concentrations of statherin in 20 mM Tris, 150 mM NaCl, pH 8, and 20 °C. The legend applies to both panels.

been long debated, urea is regarded as being able to weaken hydrophobic interactions in aqueous solution [155, 156]. Thus, that the self-association occurred both at high and low salt concentration and was hindered by urea, was interpreted as it being hydrophobically driven. To induce self-association within the model, an additional short-ranged attractive potential between neutral residues was needed, mimicking a smeared hydrophobic interaction. The strength of this potential was determined by comparing the average association number between simulations and experiments at 150 mM NaCl and 20 °C. The model was then able to capture the trends regarding protein concentration, salt concentration, and temperature. In line with the experimental findings, the complexes were shown to be more globular/spherical than the monomeric protein, see Figure 9.4a. In addition, the simulations also revealed polydispersity, as shown in Figure 9.4b. The reduction of average association number with decreased ionic strength demonstrated that electrostatic repulsion between the chains contributes to limit the growth of complexes. Substituting the phosphorylated residues with non-charged residues within the model gave larger complexes, revealing the electrostatic contribution of the phosphorylated residues. Excluding charges all together pinpointed the contribution of chain entropy in limiting the growth of complexes, which I therefore believe is the dominating factor behind the temperature effect observed in this system.

To conclude, the adjusted model successfully captured the experimentally observed trends and aided in the explanation of the observed effects in terms of a balance between different interactions and entropy. However, some limitations of the model were also encountered. First, upon inclusion of the additional attractive potential, the shape and size of the monomeric protein were no longer in agreement with SAXS data, as shown in Figure 9.5a. It might be possible to counteract this by also including an angular potential, but it would require careful balancing against the short-ranged attraction. Also, at high salt concentrations

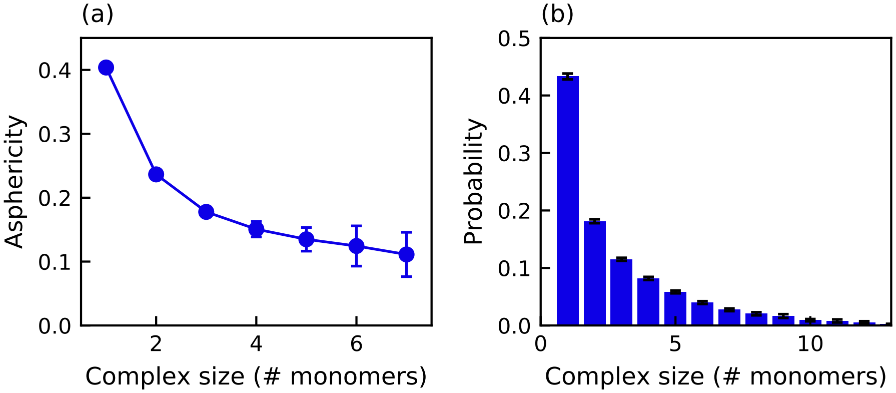


Figure 9.4: a) Asphericity of complexes of different size and b) size distribution in the simulation of 5 mg/mL statherin.

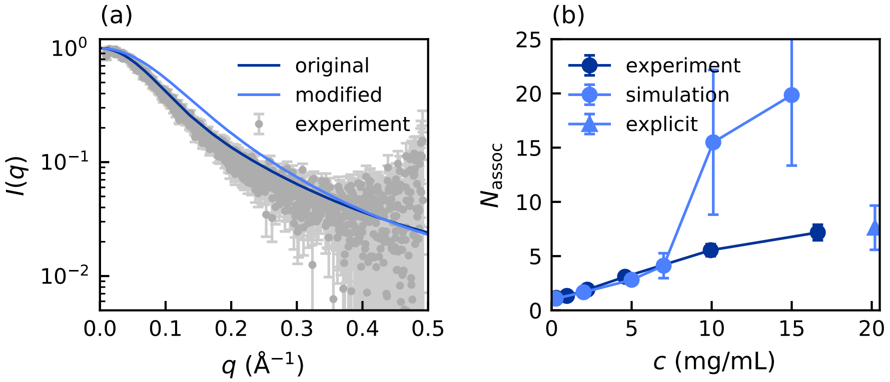


Figure 9.5: a) Form factor of statherin calculated in the original model and after inclusion of an additional short-ranged potential necessary for simulating self-association, compared to the experimentally determined form factor (collected by SEC-SAXS at pH 8 and 20 °C, with 20 mM Tris and 150 mM NaCl). b) Average association number against statherin concentration calculated from SAXS data (experimental) and determined from simulations at an ionic strength of 150 mM and 20 °C. The triangular data point is the result of a simulation using explicit salt.

the model was only applicable at low protein concentrations, as seen in Figure 9.5b. At high protein concentrations all protein chains aggregated into one large complex. This was discovered to depend on the implicit treatment of salt. With explicit salt no such breakdown was observed, which shows that the model performs better with a more accurate description of the electrostatic interactions than the extended Debye-Hückel potential. However, an explicit treatment of salt greatly increases the number of particles in the system and therefore poses larger demands on computational resources and the simulation software.

9.3 An atomistic approach to phosphorylated IDPs

The coarse-grained treatment of phosphorylated IDPs in Paper I suggested that depending on the number of phosphorylated residues and their distribution throughout the sequence, short-ranged attractive electrostatic interactions can have dramatic effects on the conformational ensemble. The discrepancies between simulations and experimental references motivated a more detailed investigation, using an atomistic approach. In addition, phosphorylation has been shown to be a versatile method for controlling protein function, as different IDPs have demonstrated varying conformational and structural response. It is therefore desirable to achieve a better understanding of phosphorylation effects.

Due to the computational expense of all-atom simulations, the 15 residue long N-terminal fragment of statherin, SN15, was chosen instead of the full protein for studying phosphorylation effects in Paper III. I selected two different force fields shown to work well for short IDPs and which had parameters for phosphorylated residues available: i) Amber ff99SB-ILDN [84] with the TIP4P-D water model [64] and the phosaa10 parameter set for phosphorylated residues [157, 158] (A99), and ii) CHARMM36m [75] with the CHARMM-modified TIP3P water model [71] (C36). Note however that the Amber parameters had been developed for a preceding force field. For experimental reference, SAXS and CD spectroscopy were performed. The force fields were shown to be in good agreement for the non-phosphorylated peptide. R_g, R_{ee} and scattering curves were in excellent agreement, and the scattering curves also matched the experimental curve, see Figure 9.6a,b. On the contrary, for the phosphorylated peptide there were large discrepancies between the force fields

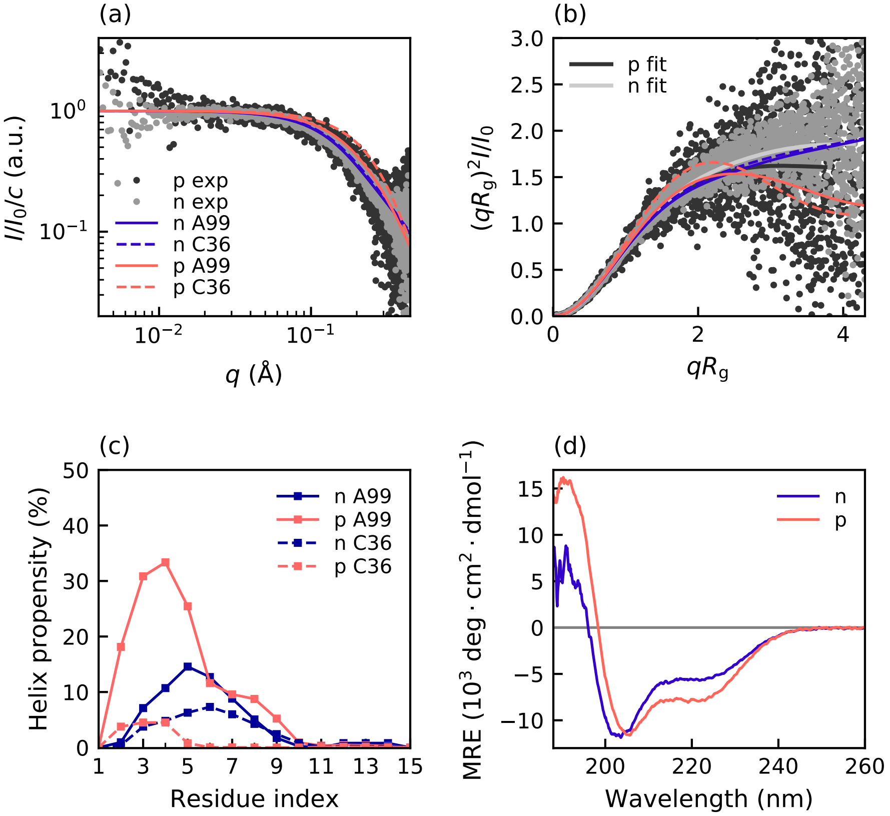


Figure 9.6: a) Form factor and b) dimensionless Kratky plot of non-phosphorylated (n) and phosphorylated (p) SN15 obtained by SAXS at 4 and 1.2 mg/mL, respectively, at 20 °C, 150 mM NaCl, 20 mM Tris, and pH 7.5, and from simulations using AMBER ff99SB-ILDN+TIP4P-D (A99) and CHARMM36m (C36). The lines "fit" correspond to the regularised curves fitted to the experimental SAXS data in the $P(r)$ determination. c) Helix propensity along the sequence for non-phosphorylated and phosphorylated SN15. d) CD spectra of non-phosphorylated and phosphorylated SN15 measured at 20 °C in 20 mM phosphate buffer with 150 mM NaF at pH 7.5, shown as the mean residue ellipticity *versus* wavelength.

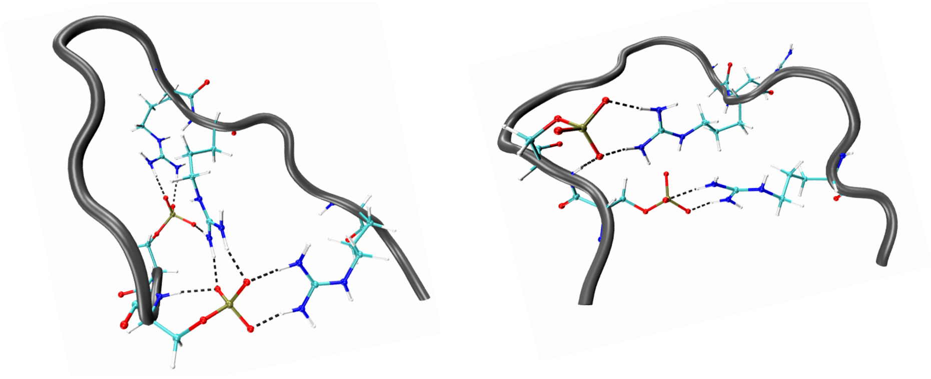


Figure 9.7: Two representative compact conformations of SN15 in CHARMM36m held together by strong salt bridges. All atoms are shown in the positively charged and phosphorylated residues. The black dashed lines represent hydrogen bonds.

regarding overall size, shape and secondary structure. C36 produced much more compact conformations, which were coupled to a higher occurrence of salt bridges between phosphorylated and positively charged residues, see Figure 9.7 for illustrative snapshots. These salt bridges also increased the content of bends in the peptide. The other main difference in secondary structure was the helical content. A substantial increase of α- and 3_{10}-helical content was observed upon phosphorylation in A99, but not in C36, as shown in Figure 9.6c. The differences in CD spectra between non-phoshorylated and phosphorylated SN15 shown in Figure 9.6d, supports an increase of α-helical structure. Both force fields gave a compaction of the peptide upon phosphorylation, however, the R_g determined from SAXS data for the non-phosphorylated and phosphorylated peptide were indistinguishable. Nonetheless, the Kratky plot indicated a small compaction upon phosphorylation, according to Figure 9.6b. Hence, a compaction in accordance with the simulations is plausible, but most probably not as large as in C36. To investigate whether the deficiencies of the force fields were general or specific to SN15, in Paper IV, the study was expanded to an additional four peptides, presented in Table 9.2.

Table 9.2: Full name, number of residues (N_{res}), phosphorylation sites (N_{ph}), positively charged residues (N_+), negatively charged residues (N_-), and net charge of the non-phosphorylated (Z_{no}) and phosphorylated variant (Z_{ph}) of the peptides studied throughout Paper III–V.

Name	Peptide	N_{res}	N_{ph}	N_+	N_-	Z_{no}	Z_{ph}
Tau1	$\text{tau}_{173-183}$	11	2	2	0	+2	-2
SN15	statherin_{1-15}	15	2	4	3	+1	-3
Tau2	$\text{tau}_{225-246}$	22	4	5	0	+5	-3
bCPP	$\beta\text{-casein}_{1-25}$	25	4	2	7	-5	-13
Stath	statherin	43	2	4	4	0	-4

C36 was shown to produce much more compact ensembles than A99 for all the phosphorylated peptides, see Figure 9.8. All peptides showed significantly higher probability of salt bridges in C36 than A99, which was the main reason behind the discrepancy between the force fields. In the 43 residue long statherin, where the phosphorylated and positively charged residues are all located within the first 13 residues, there was also another contribution. The C36 simulation contained more structures with β-strand and β-bridge formation between the middle and C-terminal end, and less structures where the protein was allowed more extended conformations. Additionally, all peptides contained a higher fraction of bends in C36, which in most cases could be linked to the salt bridges. Another noteworthy observation regarding secondary structure was that C36, in contrary to A99, did not sample any helical content at all in the N-terminal region of statherin. Although the N-terminal end of statherin is considered to be mainly irregular in water, helical propensity has been

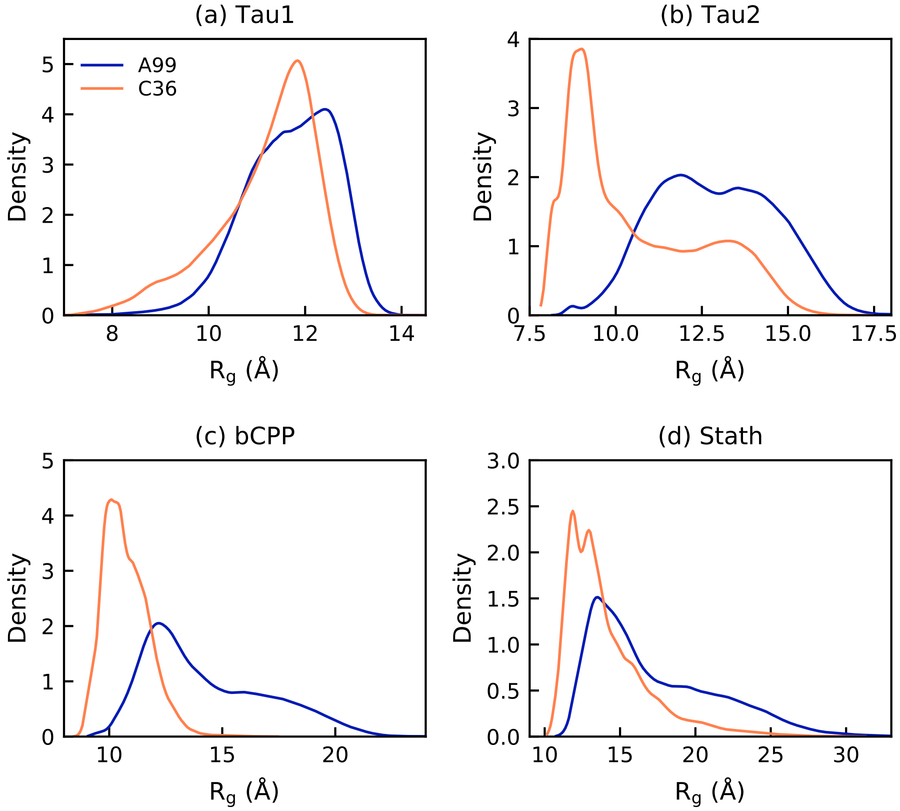


Figure 9.8: Radius of gyration distribution of a) Tau1, b) Tau2, c) bCPP and d) statherin, simulated with AMBER ff99SB-ILDN (A99) and CHARMM36m (C36).

detected in experiments [30, 159].

Noticing the large influence of salt bridges on the conformational ensemble, it was worth considering the influence of screening by addition of salt. These simulations have been performed in a salt-free environment, only with counterions to neutralise the system. So, for bCPP that showed the largest deviations between the force fields, in line with being the most charged peptide with the greatest separation of oppositely charged residues, additional simulations with 150 mM NaCl were performed. Although the probability of several salt bridges were greatly reduced in C36 when adding salt, the conformational ensemble did not change much, as was shown by comparing the R_g distributions (Figure 9.9a). In fact, the most probable conformations were still heavily influenced by salt bridges and the electrostatic interactions involving phosphorylated residues. In A99 only one salt bridge was significantly reduced, and the R_g distributions were highly similar. The calculated scattering curves were also indistinguishable in both force fields, see Figure 9.9b. Hence, the inclusion of 150 mM salt had little to no effect on the conformational ensemble, and the salt bridges were still of importance. It has been indicated that many force fields have a tendency to overestimate salt bridges [85, 141, 160, 161], hence, it is possible that both A99 and C36 overestimate the importance of salt bridges in phosphorylated IDPs. Compared to available experimental data for the shortest peptide Tau1 and the longest IDP statherin, A99 appeared as the better choice for simulating phosphorylated IDPs. However, for a better evaluation of the force fields, more experimental data is needed. Here NMR plays an important role, by being able to detect secondary secondary structure propensity for individual residues and salt bridges by scalar couplings, chemical shifts and NOEs.

In Paper v the A99 force field was used to also simulate the non-phosphorylated variants of

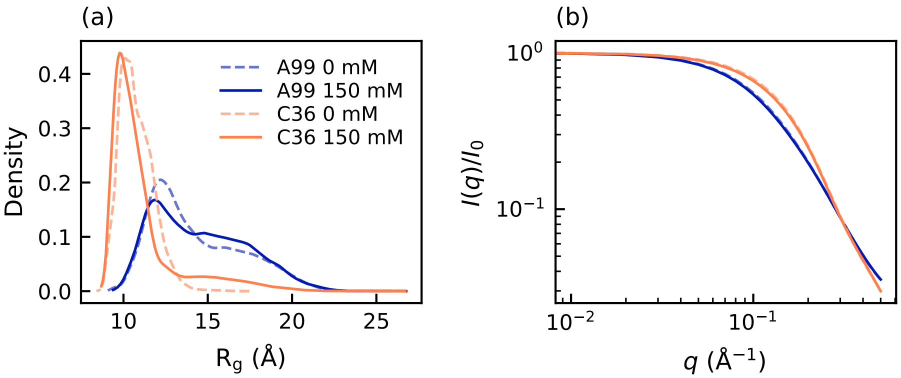


Figure 9.9: a) Radius of gyration distribution and b) calculated form factor of bCPP simulated with Amber ff99SB-ILDN (A99) or CHARMM36m (C36) in the presence of 0 or 150 mM NaCl.

the peptides, to study the conformational and structural effects induced by phosphorylation. To fully observe the electrostatic effects, the simulations were performed without additional salt. However, complementary simulations of bCPP at 150 mM demonstrated that phosphorylation effects still remained at 150 mM, although slightly diminished. Recently it was hypothesised that the global conformational changes could be predicted from the net charge of an IDP in non-phosphorylated state, such that a positively charged IDP contracts, while a neutral or negatively charge IDP expands [141]. Both Tau1 and bCPP were shown to contradict this hypothesis, see Table 9.3. In bCPP the electrostatic attraction between the arginine termini residues and the phosphorylated region drove a contraction of the peptide (see Figure 9.10), despite a local expansion of region E13–E21, containing the phosphorylation sites. Salt bridge formation between arginine/lysine and phosphorylated residues was indeed shown to be a major reason behind compaction upon phosphorylation in SN15, Tau2, and bCPP. Another contributing factor in SN15 and Tau2 was helix formation. These peptides, as well as statherin, which also exhibited increased helix propensity upon phosphorylation, all have a lysine three or four steps away from the phosphorylated residue, a pattern known to stabilise helices through salt bridge formation between the side groups [162].

In statherin, phosphorylation induced a compaction of the first 15 residues, but an overall expansion. The expansion was not caused by electrostatic repulsion, but instead explained by the preference of forming arginine-phosphoserine salt bridges over arginine–tyrosine cation–π-interaction. In non-phosphorylated statherin, arginine–tyrosine interaction caused β-sheet formation, which disappeared upon phosphorylation, when the arginine residues instead became involved in salt bridges with phosphoserine. The disruption of the β-sheet caused a global expansion. Relating back to Paper I, it is worth noticing that these effects are not captured by the coarse-grained model, since it only includes electrostatic effects between charged residues. In fact, the coarse-grained model provides a small decrease in R_g upon phosphorylation, originating from the compaction of the N-terminal region where the phosphorylated residues reside.

To conclude, the studies performed in Paper III-V showed that phosphorylation induces changes in both overall dimensions and structural content, and that salt bridge formation

Table 9.3: Net charge of the non-phosphorylated peptide and mean radius of gyration (R_g) and end-to-end distance (R_{ee}) of the non-phosphorylated (n) and phosphorylated (p) variants.

		R_g (Å)		R_{ee} (Å)	
Peptide	Net charge	n	p	n	p
Tau1	+2	9.3 ± 0.1	9.8 ± 0.1	27.4 ± 0.6	28.9 ± 0.2
SN15	+1	10.0 ± 0.1	9.0 ± 0.1	25.4 ± 0.9	23.0 ± 0.3
Tau2	+5	14.6 ± 0.2	12.9 ± 0.3	38.3 ± 0.9	32.7 ± 1.7
bCPP	-5	15.3 ± 0.3	14.3 ± 0.3	38.0 ± 0.8	30.9 ± 1.5
Stath	0	15.6 ± 0.4	17.3 ± 0.9	33.0 ± 0.4	40.5 ± 1.7

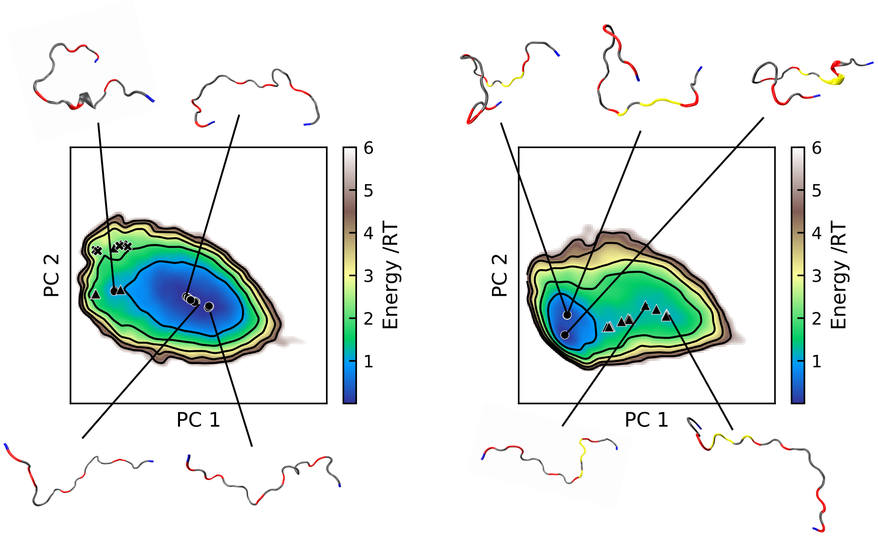


Figure 9.10: Energy landscapes with conformations in selected minima of bCPP for non-phosphorylated (left) and phosphorylated (right) bCPP. The energy landscapes were constructed using the first two components from principal component analysis, using the same basis set for both variants. Hence, they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$, ✖: $\leq 3RT$. In the conformations positively charged residues are shown in blue, negatively charged residues in red and phosphorylated residues in yellow.

is an important contributor to this. Vast over-stabilisation of salt bridges was shown to have large effects on the global dimensions, demonstrating the need for revised force fields. Also at 150 mM salt did salt bridges between phosphorylated and positively charged residues influence the conformational ensemble. It was shown that only considering net charge is not enough for predicting the outcome of phosphorylation, and that also non-charged residues can be of importance. Atomistic simulations show great potential in providing deeper knowledge regarding the effect of phosphorylation, however, more experimental studies at both global and local length-scales are required for further revision and validation of force fields.

9.4 Conclusions and outlook

The overall objective of this thesis has been to investigate the conformational ensembles exhibited by IDPs in solution, to explore the dependence on sequence, especially the impact of phosphorylated residues. Due to the conformational polydispersity exhibited by IDPs, it is challenging to extract detailed information from experiments, but combining different experimental and computational techniques has proven to be a fruitful approach. Since a computational approach is dependent on appropriate models, a significant part of the work has been focused on investigating how models and force fields perform.

One property characterising a great model is it being as simple as possible, but still describing the phenomenon of interest. In this way, it can act as an explanatory tool. The coarse-grained ”one bead per residue model” relying on excluded volume, electrostatic interactions and an approximate van der Waals interaction was shown to reproduce R_g for a range of different IDPs under dilute conditions, implying that many IDPs can be thought of as self-avoiding random walks influenced by electrostatic interactions. From this model, a basic understanding of how chain length, charge distribution and salt concentration affects the conformational ensemble can be achieved. Furthermore, with the addition of a hydrophobic interaction, the model was shown to qualitatively describe the self-association process of statherin and provided a deeper understanding of the balance of interactions. This demonstrates that the model is applicable also in larger and more complex systems, where coarse-grained approaches are currently the only feasible option considering the computational expense versus resources. Other adaptations of the original model have also been applied to studies of crowding [123, 163] and zinc-initiated oligomerisation [164], showcasing the potential and adaptability of this model within the field of IDP research. However, all models come with limitations. Here it was shown that the model in current form could not simultaneously provide a good representation of both size and level of stiffness for the proline-rich proteins and that the size of the highly phosphorylated IDPs was underestimated. Since IDPs are a very diverse group of proteins, it is by no means surprising that not all IDPs can be described by this model. For the phosphorylated proteins, better agreement was achieved with a reduced charge of the phosphorylated residues. It is therefore of interest to further explore whether this is due to an overestimation of electrostatic interactions in the model, ill-matching of the experimental conditions or if a fixed charge of $-2e$ is a poor representation of the charge state of phosphorylated residues at physiological pH. Also, in the simulations of self-association, the implicit treatment of salt caused the model to break down at higher protein concentrations. While an explicit treatment of salt provides better results, it comes with a larger computational cost and limits to the accessible system size.

Regarding the effects of phosphorylation, this problem required a more detailed model. Atomistic simulations were shown to detect changes in global compaction and secondary structure, and relate them to interactions between specific residues. Especially salt bridges between phosphorylated and positively charged residues were shown to have major impact on the conformational ensemble, which highlighted the importance of having force fields that accurately estimate the strength of salt bridges. Other force field deficiencies regarding secondary structure were also detected. In the continued strive for understanding the implications of phosphorylation of IDPs, it is therefore important to revise force fields, and to especially consider the strength of salt bridges involving phosphorylated residues. Therefore, the collection of more experimental data suitable for use as benchmarking is also required, which extends beyond the techniques applied in this work. NMR was mentioned as an example, which has the advantage that scalar couplings and chemical shifts can be calculated from simulations, which facilitates comparison. The interplay between arginines, tyrosines and phosphorylated residues implied by the atomistic simulations of statherin is of specific interest to explore further. In addition, a systematic investigation varying the number of phosphorylated residues and their position in relation to positively charged residues in a controlled manner is suggested for gaining a better understanding of underlaying factors controlling the outcome of phosphorylation.

While this thesis has been focused on the relation between sequence and structure, an area where much is yet to be explored, the link to function is equally important to consider. Since the functionality often involves interaction with binding partners or surfaces, there is a requirement for computational models to handle such situations. Also in this context can statherin be used as a model protein, as binding to hydroxyapatite has been shown to induce more helix formation in the N-terminal end [165, 166] and expose a bacterial binding site in the C-terminal tail [166, 167].

As a final remark, one of the greatest lessons I have learned during these years of research is that it is not at all straightforward to compare experimental and simulation data and draw correct conclusions from it. Here I see great advantages of having practical experience of both parts, as it provides better comprehension of what can affect the data and what is actually compared.

References

[1] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, "Classification of intrinsically disordered regions and proteins," *Chem. Rev.*, vol. 114, no. 13, pp. 6589–6631, 2014.

[2] C. J. Oldfield and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions," *Annu. Rev. Biochem.*, vol. 83, no. 1, pp. 553–584, 2014.

[3] P. E. Wright and H. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *J. Mol. Biol.*, vol. 293, no. 2, pp. 321 – 331, 1999.

[4] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovićá, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[5] V. N. Uversky and A. K. Dunker, "Understanding protein non-folding," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1804, no. 6, pp. 1231 – 1264, 2010.

[6] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*. New York, USA: W. H. Freeman and Company, international 7th ed., 2011.

[7] Y. Mansiaux, A. P. Joseph, J.-C. Gelly, and A. G. de Brevern, "Assignment of polyproline ii conformation and analysis of sequence – structure relationship," *PLOS ONE*, vol. 6, pp. 1–15, 03 2011.

[8] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[9] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, no. 1, pp. 38–48, 2001.

[10] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[11] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Inform.*, vol. 11, pp. 161–171, 2000.

[12] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca, E. Garner, S. Guilliot, and A. Dunker, "Thousands of proteins likely to have long disordered regions," *Pac. Symp. Biocomput.*, vol. 3, pp. 437–448, 1998.

[13] J. Ward, J. Sodhi, L. McGuffin, B. Buxton, and D. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J. Mol. Biol.*, vol. 337, no. 3, pp. 635–645, 2004.

[14] B. Xue, A. K. Dunker, and V. N. Uversky, "Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life," *J. Biomol. Struct. Dyn.*, vol. 30, no. 2, pp. 137–149, 2012.

[15] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.*, vol. 6, pp. 197–208, 2005.

[16] P. Tompa, "Intrinsically disordered proteins: a 10-year recap," *Trends Biochem. Sci.*, vol. 37, no. 12, pp. 509 – 516, 2012.

[17] J. Liu, J. R. Faeder, and C. J. Camacho, "Toward a quantitative theory of intrinsically disordered proteins and their function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 47, pp. 19819–19823, 2009.

[18] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Curr. Opin. Struct. Biol.*, vol. 19, no. 1, pp. 31–38, 2009.

[19] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: Introducing the d2 concept," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 215–246, 2008.

[20] V. N. Uversky, V. Davé, L. M. Iakoucheva, P. Malaney, S. J. Metallo, R. R. Pathak, and A. C. Joerger, "Pathological unfoldomics of uncontrolled chaos: Intrinsically disordered proteins and human diseases," *Chem. Rev.*, vol. 114, no. 13, pp. 6844–6879, 2014.

[21] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *J. Mol. Graphics Modell.*, vol. 19, no. 1, pp. 26–59, 2001.

[22] V. N. Uversky, "Unusual biophysics of intrinsically disordered proteins," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1834, no. 5, pp. 932–951, 2013.

[23] R. K. Das, K. M. Ruff, and R. V. Pappu, "Relating sequence encoded information to form and function of intrinsically disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 32, pp. 102–112, 2015. New constructs and expression of proteins / Sequences and topology.

[24] R. K. Das and R. V. Pappu, "Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 33, pp. 13392–13397, 2013.

[25] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker, "The importance of intrinsic disorder for protein phosphorylation," *Nucleic Acids Res.*, vol. 32, pp. 1037–1049, 02 2004.

[26] J. Gao and D. Xu, *Biocomputing 2012*, ch. Correlation Between Posttranslational Modification and Intrinsic Disorder in Protein, pp. 94–103. World Scientific Publishing Co. Pte. Ltd., 2012.

[27] L. N. Johnson and R. J. Lewis, "Structural basis for control by phosphorylation," *Chem. Rev.*, vol. 101, no. 8, pp. 2209–2242, 2001.

[28] C. X. Gong and K. Iqbal, "Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for alzheimer disease," *Curr. Med. Chem.*, vol. 15, no. 23, pp. 2321–2328, 2008.

[29] C. G. De Kruif and C. Holt, *Casein Micelle Structure, Functions and Interactions*, pp. 233–276. Boston, MA: Springer US, 2003.

[30] P. A. Raj, M. Johnsson, M. J. Levine, and G. H. Nancollas, "Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization.," *J. Biol. Chem.*, vol. 267, no. 9, pp. 5968–76, 1992.

[31] K. Makrodimitris, D. L. Masica, E. T. Kim, and J. J. Gray, "Structure prediction of protein–solid surface interactions reveals a molecular recognition motif of statherin for hydroxyapatite," *J. Am. Chem. Soc.*, vol. 129, no. 44, pp. 13713–13722, 2007.

[32] J. A. Loo, W. Yan, P. Ramachandran, and D. T. Wong, "Comparative human salivary and plasma proteomes," *J. Dent. Res.*, vol. 89, no. 10, pp. 1016–1023, 2010.

[33] M. Edgar, C. Dawes, and D. O'Mullane, eds., *Saliva and Oral Health*. London, UK: British Dental Association, 3rd ed., 2004.

[34] W. Siqueira, W. Custodio, and E. McDonald, “New insights into the composition and functions of the acquired enamel pellicle,” *J. Dent. Res.*, vol. 91, no. 12, pp. 1110–1118, 2012.

[35] M. J. Levine, “Development of artificial salivas,” *Crit. Rev. Oral Biol. Med.*, vol. 4, no. 3, pp. 279–286, 1993.

[36] E. Moreno and R. Zahradnik, “Demineralization and remineralization of dental enamel,” *J. Dent. Res.*, vol. 58, no. 2_suppl, pp. 896–903, 1979.

[37] D. Hay, D. Smith, S. Schluckebier, and E. Moreno, “Basic biological sciences relationship between concentration of human salivary statherin and inhibition of calcium phosphate precipitation in stimulated human parotid saliva,” *J. Dent. Res.*, vol. 63, no. 6, pp. 857–863, 1984.

[38] M. A. Buzalaf, A. R. Hannas, and M. T. Kato, “Saliva and dental erosion,” *J. Appl. Oral Sci.*, vol. 20, no. 5, pp. 493–502, 2012.

[39] W. H. Douglas, E. S. Reeh, N. Ramasubbu, P. A. Raj, K. K. Bhandary, and M. J. Levine, “Statherin: A major boundary lubricant of human saliva,” *Biochem. Biophys. Res. Commun.*, vol. 180, no. 1, pp. 91 – 97, 1991.

[40] R. J. Gibbons and D. I. Hay, “Human salivary acidic proline-rich proteins and statherin promote the attachment of actinomyces viscosus LY7 to apatitic surfaces.,” *Infect. Immun.*, vol. 56, no. 2, pp. 439–445, 1988.

[41] A. Amano, K. Kataoka, P. A. Raj, R. J. Genco, and S. Shizukuishi, “Binding sites of salivary statherin for porphyromonas gingivalis recombinant fimbrillin,” *Infect. Immun.*, vol. 64, no. 10, pp. 4249–4254, 1996.

[42] H. Nagata, A. Sharma, H. T. Sojar, A. Amano, M. J. Levine, and R. J. Genco, “Role of the carboxyl-terminal region of porphyromonas gingivalis fimbrillin in binding to salivary proteins,” *Infect. Immun.*, vol. 65, no. 2, pp. 422–427, 1997.

[43] D. H. Schlesinger and D. I. Hay, “Complete covalent structure of statherin, a tyrosine-rich acidic peptide which inhibits calcium phosphate precipitation from human parotid saliva,” *J. Biol. Chem.*, vol. 252, no. 5, pp. 1689–1695, 1977.

[44] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, 1982.

[45] C. Holt, “Unfolded phosphopolypeptides enable soft and hard tissues to coexist in the same organism with relative ease,” *Curr. Opin. Struct. Biol.*, vol. 23, no. 3, pp. 420–425, 2013. New contructs and expressions of proteins / Sequences and topology.

[46] Y. Lin, S. L. Currie, and M. K. Rosen, "Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs," *J. Biol. Chem.*, vol. 292, no. 46, pp. 19110–19120, 2017.

[47] C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, A. A. Yunus, D. Liu, R. V. Pappu, and M. K. Rosen, "Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein," *Mol. Cell*, vol. 63, no. 1, pp. 72–85, 2016.

[48] E. Rieloff, M. D. Tully, and M. Skepö, "Assessing the intricate balance of intermolecular interactions upon self-association of intrinsically disordered proteins," *J. Mol. Biol.*, vol. 431, no. 3, pp. 511–523, 2019.

[49] J. N. Israelachvili, *Intermolecular and Surface Forces*. Oxford, UK: Academic Press, Elsevier, 3rd ed., 2011.

[50] M. T. A. Evans, M. C. Phillips, and M. N. Jones, "The conformation and aggregation of bovine β-casein a. II. Thermodynamics of thermal association and the effects of changes in polar and apolar interactions on micellization," *Biopolymers*, vol. 18, no. 5, pp. 1123–1140, 1979.

[51] K. Takase, R. Niki, and S. Arima, "A sedimentation equilibrium study of the temperature-dependent association of bovine β-casein," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 622, no. 1, pp. 1–8, 1980.

[52] J. O'Connell, V. Grinberg, and C. de Kruif, "Association behavior of β-casein," *J. Colloid Interface Sci.*, vol. 258, no. 1, pp. 33–39, 2003.

[53] I. Portnaya, U. Cogan, Y. D. Livney, O. Ramon, K. Shimoni, M. Rosenberg, and D. Danino, "Micellization of bovine β-casein studied by isothermal titration microcalorimetry and cryogenic transmission electron microscopy," *J. Agric. Food Chem.*, vol. 54, no. 15, pp. 5555–5561, 2006.

[54] C. Moitzi, I. Portnaya, O. Glatter, O. Ramon, and D. Danino, "Effect of temperature on self-assembly of bovine β-casein above and below isoelectric pH. Structural analysis by cryogenic-transmission electron microscopy and small-angle x-ray scattering," *Langmuir*, vol. 24, no. 7, pp. 3020–3029, 2008.

[55] D. Chandler, "Hydrophobicity: Two faces of water," *Nature*, vol. 417, no. 491, pp. 493–502, 2002.

[56] T. L. Hill, *An Introduction to Statistical Thermodynamics*. Reading, MA, USA: Addison-Wesley Publishing Company, 2nd ed., 1962.

[57] C. Cragnell, D. Durand, B. Cabane, and M. Skepö, "Coarse-grained modeling of the intrinsically disordered protein histatin 5 in solution: Monte carlo simulations in combination with saxs," *Proteins*, vol. 84, no. 6, pp. 777–791, 2016.

[58] H. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Comput. Phys. Commun.*, vol. 91, no. 1, pp. 43–56, 1995.

[59] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, 2008.

[60] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, pp. 845–854, 02 2013.

[61] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, "Tackling exascale software challenges in molecular dynamics simulations with gromacs," in *Solving Software Challenges for Exascale* (S. Markidis and E. Laure, eds.), (Cham), pp. 3–27, Springer International Publishing, 2015.

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

[63] G. P. Moss, "Basic terminology of stereochemistry (IUPAC recommendations 1996)," *Pure Appl. Chem.*, vol. 68, no. 12, pp. 2193–2222, 1996.

[64] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water dispersion interactions strongly influence simulated structural properties of disordered protein states," *J. Phys. Chem. B*, vol. 119, no. 16, pp. 5113–5123, 2015.

[65] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, "Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment," *J. Chem. Theory Comput.*, vol. 11, no. 11, pp. 5513–5524, 2015.

[66] J. Henriques and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: On the accuracy of the TIP4P-D water model and the representativeness of protein disorder models," *J. Chem. Theory Comput.*, vol. 12, no. 7, pp. 3407–3415, 2016.

[67] A. V. Onufriev and S. Izadi, "Water models for biomolecular simulations," *WIREs Comput. Mol. Sci.*, vol. 8, no. 2, p. e1347, 2018.

[68] W. L. Jorgensen, "Transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water," *J. Am. Chem. Soc.*, vol. 103, pp. 335–340, 1981.

[69] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[70] S. R. Durell, B. R. Brooks, and A. Ben-Naim, "Solvent-induced forces between two hydrophilic groups," *J. Phys. Chem.*, vol. 98, pp. 2198–2202, 1994.

[71] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[72] J. L. F. Abascal and C. Vega, "A general purpose model for the condensed phases of water: TIP4P/2005," *J. Chem. Phys.*, vol. 123, no. 23, p. 234505, 2005.

[73] O. Guvench and A. D. MacKerell, *Comparison of Protein Force Fields for Molecular Dynamics Simulations*, pp. 63–88. Totowa, NJ: Humana Press, 2008.

[74] S. Boonstra, P. R. Onck, and E. van der Giessen, "CHARMM TIP3P water model suppresses peptide folding by solvating the unfolded state," *J. Phys. Chem. B*, vol. 120, no. 15, pp. 3692–3698, 2016.

[75] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell Jr, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nat. Methods.*, vol. 14, no. 1, pp. 71–73, 2017.

[76] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?," *Biophys. J.*, vol. 95, no. 1, pp. L07–L09, 2008.

[77] W. Wang, W. Ye, C. Jiang, R. Luo, and H.-F. Chen, "New force field on modeling intrinsically disordered proteins," *Chem. Biol. Drug. Des.*, vol. 84, no. 3, pp. 253–269, 2014.

[78] Y. Zhang, H. Liu, S. Yang, R. Luo, and H.-F. Chen, "Well-balanced force field ff03CMAP for folded and disordered proteins," *J. Chem. Theory Comput.*, vol. 15, no. 12, pp. 6769–6780, 2019.

[79] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular

dynamics simulations," *Curr. Opin. Struct. Biol.*, vol. 24, pp. 98–105, 2014. Folding and binding / Nucleic acids and their protein complexes.

[80] J. Henriques, C. Cragnell, and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment," *J. Chem. Theory Comput.*, vol. 11, no. 7, pp. 3420–3431, 2015.

[81] A. D. Mackerell Jr., M. Feig, and C. L. Brooks III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," *J. Comput. Chem.*, vol. 25, no. 11, pp. 1400–1415, 2004.

[82] R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides," *J. Phys. Chem. B*, vol. 113, no. 26, pp. 9004–9015, 2009.

[83] R. B. Best and J. Mittal, "Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse," *J. Phys. Chem. B*, vol. 114, no. 46, pp. 14916–14923, 2010.

[84] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins*, vol. 78, no. 8, pp. 1950–1958, 2010.

[85] S. Piana, K. Lindorff-Larsen, and D. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," *Biophys. J.*, vol. 100, no. 9, pp. L47–L49, 2011.

[86] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3257–3273, 2012.

[87] R. B. Best, W. Zheng, and J. Mittal, "Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association," *J. Chem. Theory Comput.*, vol. 10, no. 11, pp. 5113–5124, 2014.

[88] F. Jiang, C.-Y. Zhou, and Y.-D. Wu, "Residue-specific force field based on the protein coil library. RSFF1: Modification of OPLS-AA/L," *J. Phys. Chem. B*, vol. 118, no. 25, pp. 6983–6998, 2014.

[89] C.-Y. Zhou, F. Jiang, and Y.-D. Wu, "Residue-specific force field based on protein coil library. rsff2: Modification of amber ff99sb," *J. Phys. Chem. B*, vol. 119, no. 3, pp. 1035–1047, 2015.

[90] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, 2015.

[91] D. Song, R. Luo, and H.-F. Chen, "The idp-specific force field ff14idpsff improves the conformer sampling of intrinsically disordered proteins," *J. Chem. Inf. Model.*, vol. 57, no. 5, pp. 1166–1178, 2017.

[92] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 21, pp. E4758–E4766, 2018.

[93] H. Liu, D. Song, H. Lu, R. Luo, and H.-F. Chen, "Intrinsically disordered protein-specific force field CHARMM36IDPSFF," *Chem. Biol. Drug. Des.*, vol. 92, no. 4, pp. 1722–1735, 2018.

[94] H. Liu, D. Song, Y. Zhang, S. Yang, R. Luo, and H.-F. Chen, "Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins," *Phys. Chem. Chem. Phys.*, vol. 21, pp. 21918–21931, 2019.

[95] S. Yang, H. Liu, Y. Zhang, H. Lu, and H. Chen, "Residue-specific force field improving the sample of intrinsically disordered proteins and folded proteins," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4793–4805, 2019.

[96] J. Mu, H. Liu, J. Zhang, R. Luo, and H.-F. Chen, "Recent force field strategies for intrinsically disordered proteins," *J. Chem. Inf. Model.*, vol. 61, no. 3, pp. 1037–1047, 2021.

[97] J. Huang and A. D. MacKerell, "Force field development and simulations of intrinsically disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 48, pp. 40–48, 2018. Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective.

[98] S.-H. Chong, P. Chatterjee, and S. Ham, "Computer simulations of intrinsically disordered proteins," *Annu. Rev. Phys. Chem.*, vol. 68, no. 1, pp. 117–134, 2017.

[99] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[100] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*. San Diego, CA, USA: Academic Press, 2nd ed., 2002.

[101] J. Reščič and P. Linse, "MOLSIM: A modular molecular simulation software," *J. Comput. Chem.*, vol. 36, no. 16, pp. 1259–1274, 2015.

[102] M. Allen and D. Tildesley, *Computer Simulation of Liquids*. Oxford University Press, 1989.

[103] M. Abraham, B. Hess, D. van der Spoel, and E. Lindahl, *GROMACS Reference Manual version 2018.4*. The GROMACS development teams, www.gromacs.org.

[104] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[105] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, “Long-time-step molecular dynamics through hydrogen mass repartitioning,” *J. Chem. Theory Comput.*, vol. 11, no. 4, pp. 1864–1874, 2015.

[106] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, “Lincs: A linear constraint solver for molecular simulations,” *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997.

[107] G. Bussi, D. Donadio, and M. Parrinello, “Canonical sampling through velocity rescaling,” *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, 2007.

[108] M. Parrinello and A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method,” *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, 1981.

[109] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, 1984.

[110] D. Svergun, C. Barberato, and M. H. J. Koch, “Crysol– a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates,” *J. Appl. Crystallogr.*, vol. 28, no. 6, pp. 768–773, 1995.

[111] J. Henriques, L. Arleth, K. Lindorff-Larsen, and M. Skepö, “On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations,” *J. Mol. Biol.*, vol. 430, no. 16, pp. 2521–2539, 2018. Intrinsically Disordered Proteins.

[112] P. Chen and J. Hub, “Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data,” *Biophys. J.*, vol. 107, no. 2, pp. 435–447, 2014.

[113] Y. Hayashi, M. Ullner, and P. Linse, “Complex formation in solutions of oppositely charged polyelectrolytes at different polyion compositions and salt content,” *J. Phys. Chem. B*, vol. 107, no. 32, pp. 8198–8207, 2003.

[114] H. Arkın and W. Janke, "Gyration tensor based analysis of the shapes of polymer chains in an attractive spherical cage," *J. Chem. Phys.*, vol. 138, no. 5, p. 054904, 2013.

[115] M. Kenward and M. D. Whitmore, "A systematic monte carlo study of self-assembling amphiphiles in solution," *J. Chem. Phys.*, vol. 116, no. 8, pp. 3455–3470, 2002.

[116] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[117] R. Chebrek, S. Leonard, A. G. de Brevern, and J.-C. Gelly, "PolyprOnline: polyproline helix II and secondary structure assignment database," *Database*, vol. 2014, 11 2014. bau102.

[118] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins*, vol. 23, no. 4, pp. 566–579, 1995.

[119] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *J. Mol. Graph.*, vol. 14, pp. 33–38, 1996.

[120] Y. Zhang and C. Sagui, "Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI," *J. Mol. Graph. Model.*, vol. 55, pp. 72–84, 2015.

[121] L. Mavridis and R. W. Janes, "PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates," *Bioinformatics*, vol. 33, pp. 56–63, 09 2016.

[122] G. Nagy, M. Igaev, N. C. Jones, S. V. Hoffmann, and H. Grubmüller, "Sesca: Predicting circular dichroism spectra from protein molecular structures," *J. Chem. Theory Comput.*, vol. 15, no. 9, pp. 5087–5102, 2019.

[123] E. Fagerberg, L. K. Månsson, S. Lenton, and M. Skepö, "The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions," *J. Phys. Chem. B*, vol. 124, no. 52, pp. 11843–11853, 2020.

[124] S. Jephthah, F. Pesce, K. Lindorff-Larsen, and M. Skepö, "Force field effects in simulations of flexible peptides with varying polyproline II propensity," *J. Chem. Theory Comput.*, 2021.

[125] P. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H. Ogasawara, L. Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson, "The structure of the first coordination shell in liquid water," *Science*, vol. 304, no. 5673, pp. 995–999, 2004.

[126] S. R. R. Campos and A. M. Baptista, "Conformational analysis in a multidimensional energy landscape: Study of an arginylglutamate repeat," *J. Phys. Chem. B*, vol. 113, no. 49, pp. 15989–16001, 2009.

[127] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016.

[128] A. Grossfield and D. M. Zuckerman, "Chapter 2 quantifying uncertainty and sampling quality in biomolecular simulations," vol. 5 of *Annual Reports in Computational Chemistry*, pp. 23–48, Elsevier, 2009.

[129] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, and D. M. Zuckerman, "Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]," *Living Journal of Computational Molecular Science*, vol. 1, p. 5067, Oct. 2018.

[130] B. J. H. Kuipers and H. Gruppen, "Prediction of molar extinction coefficients of proteins and peptides using uv absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography–mass spectrometry analysis," *J. Agric. Food Chem.*, vol. 55, no. 14, pp. 5445–5451, 2007.

[131] E. Mihalyi, "Numerical values of the absorbances of the aromatic amino acids in acid, neutral, and alkaline solutions," *J. Chem. Eng. Data*, vol. 13, no. 2, pp. 179–182, 1968.

[132] D. I. Svergun, M. H. J. Koch, P. A. Timmins, and R. P. May, *Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford, UK: Oxford University Press, 1st ed., 2013.

[133] J. Pérez and P. Vachette, *A Successful Combination: Coupling SE-HPLC with SAXS*, pp. 183–199. Singapore: Springer Singapore, 2017.

[134] Guinier, André, "La diffraction des rayons x aux très petits angles : application à l'étude de phénomènes ultramicroscopiques," *Ann. Phys.*, vol. 11, no. 12, pp. 161–237, 1939.

[135] V. Receveur-Bréchot and D. Durand, "How random are intrinsically disordered proteins? a small angle scattering perspective," *Curr. Protein Pept. Sci.*, vol. 13, pp. 55–75, 2012.

[136] D. Orthaber, A. Bergmann, and O. Glatter, "SAXS experiments on absolute scale with Kratky systems using water as a secondary standard," *J. Appl. Crystallogr.*, vol. 33, pp. 218–225, Apr 2000.

[137] D. Durand, C. Vivès, D. Cannella, J. Pérez, E. Pebay-Peyroula, P. Vachette, and F. Fieschi, "NADPH oxidase activator p67$^{\mathrm{phox}}$ behaves in solution as a multidomain protein with semi-flexible linkers," *J. Struct. Biol.*, vol. 169, no. 1, pp. 45 – 53, 2010.

[138] O. Glatter, "Data evaluation in small angle scattering: calculation of the radial electron density distribution by means of indirect fourier transformation," *Acta Phys. Austriaca*, vol. 47, no. 1-2, pp. 83–102, 1977.

[139] D. I. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *J. Appl. Crystallogr.*, vol. 25, no. 4, pp. 495–503, 1992.

[140] D. A. Jacques and J. Trewhella, "Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls," *Protein Sci.*, vol. 19, no. 4, pp. 642–657, 2010.

[141] F. Jin and F. Gräter, "How multisite phosphorylation impacts the conformations of intrinsically disordered proteins," *PLoS Comput. Biol.*, vol. 17, no. 5, p. e1008939, 2021.

[142] A. Miles and B. Wallace, "Chapter 6 - circular dichroism spectroscopy for protein characterization: Biopharmaceutical applications," in *Biophysical Characterization of Proteins in Developing Biopharmaceuticals* (D. J. Houde and S. A. Berkowitz, eds.), pp. 109 – 137, Amsterdam: Elsevier, 2015.

[143] S. M. Kelly, T. J. Jess, and N. C. Price, "How to study proteins by circular dichroism," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1751, no. 2, pp. 119 – 139, 2005.

[144] L. Whitmore, A. J. Miles, L. Mavridis, R. W. Janes, and B. Wallace, "PCDDB: new developments at the Protein Circular Dichroism Data Bank," *Nucleic Acids Res.*, vol. 45, pp. D303–D307, 09 2016.

[145] A. Abdul-Gader, A. J. Miles, and B. A. Wallace, "A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy," *Bioinformatics*, vol. 27, pp. 1630–1636, 04 2011.

[146] J. L. S. Lopes, A. J. Miles, L. Whitmore, and B. A. Wallace, "Distinct circular dichroism spectroscopic signatures of polyproline II and unordered secondary structures: Applications in secondary structure analyses," *Protein Sci.*, vol. 23, no. 12, pp. 1765–1772, 2014.

[147] J. Tolchard, S. J. Walpole, A. J. Miles, R. Maytum, L. A. Eaglen, T. Hackstadt, B. A. Wallace, and T. M. A. Blumenschein, "The intrinsically disordered tarp protein from chlamydia binds actin with a partially preformed helix," *Sci. Rep.*, vol. 8, no. 1, p. 1960, 2018.

[148] N. Sreerama and R. W. Woody, "Computation and analysis of protein circular dichroism spectra," in *Numerical Computer Methods, Part D*, vol. 383 of *Methods in Enzymology*, pp. 318 – 351, Academic Press, 2004.

[149] B. Schuler, A. Soranno, H. Hofmann, and D. Nettels, "Single-molecule fret spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins," *Annu. Rev. Biophys.*, vol. 45, no. 1, pp. 207–231, 2016.

[150] J. A. Riback, M. A. Bowman, A. M. Zmyslowski, K. W. Plaxco, P. L. Clark, and T. R. Sosnick, "Commonly used fret fluorophores promote collapse of an otherwise disordered protein," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 18, pp. 8889–8894, 2019.

[151] M. Carballo-Pacheco and B. Strodel, "Comparison of force fields for alzheimer's a : A case study for intrinsically disordered proteins," *Protein Sci.*, vol. 26, no. 2, pp. 174–185, 2017.

[152] G. H. Zerze, W. Zheng, R. B. Best, and J. Mittal, "Evolution of all-atom protein force fields to improve local and global properties," *J. Phys. Chem. Lett.*, vol. 10, no. 9, pp. 2227–2234, 2019.

[153] E. W. Martin, A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, and T. Mittag, "Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation," *J. Am. Chem. Soc.*, vol. 138, no. 47, pp. 15323–15335, 2016.

[154] E. Bienkiewicz and K. Lumb, "Random-coil chemical shifts of phosphorylated amino acids," *J. Biomol. NMR*, vol. 15, no. 3, pp. 203–206, 1999.

[155] R. Zangi, R. Zhou, and B. J. Berne, "Urea's action on hydrophobic interactions," *J. Am. Chem. Soc.*, vol. 131, no. 4, pp. 1535–1541, 2009.

[156] L. Costantino, G. D'Errico, P. Roscigno, and V. Vitagliano, "Effect of urea and alkylureas on micelle formation by a nonionic surfactant with short hydrophobic tail at 25 °c," *J. Phys. Chem. B*, vol. 104, no. 31, pp. 7326–7333, 2000.

[157] N. Homeyer, A. H. C. Horn, H. Lanig, and H. Sticht, "Amber force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine," *J. Mol. Model.*, vol. 12, pp. 281–289, Feb 2006.

[158] T. Steinbrecher, J. Latzer, and D. A. Case, "Revised amber parameters for bioorganic phosphates," *J. Chem. Theory Comput.*, vol. 8, no. 11, pp. 4405–4412, 2012.

[159] G. A. Naganagowda, T. L. Gururaja, and M. J. Levine, "Delineation of conformational preferences in human salivary statherin by 1h, 31p nmr and cd studies: Sequential assignment and structure-function correlations," *J. Biomol. Struct. Dyn.*, vol. 16, no. 1, pp. 91–107, 1998.

[160] K. T. Debiec, A. M. Gronenborn, and L. T. Chong, "Evaluating the strength of salt bridges: A comparison of current biomolecular force fields," *J. Phys. Chem. B*, vol. 118, no. 24, pp. 6561–6569, 2014.

[161] M. C. Ahmed, E. Papaleo, and K. Lindorff-Larsen, "How well do force fields capture the strength of salt bridges in proteins?," *PeerJ*, vol. 6, p. e4967, June 2018.

[162] N. Errington and A. J. Doig, "A phosphoserine–lysine salt bridge within an α-helical peptide, the strongest α-helix side-chain interaction measured to date," *Biochemistry*, vol. 44, no. 20, pp. 7553–7558, 2005.

[163] E. Fagerberg, S. Lenton, and M. Skepö, "Evaluating models of varying complexity of crowded intrinsically disordered protein solutions against SAXS," *J. Chem. Theory Comput.*, vol. 15, no. 12, pp. 6968–6983, 2019.

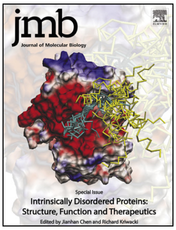
[164] C. Cragnell, L. Staby, S. Lenton, B. B. Kragelund, and M. Skepö, "Dynamical oligomerisation of histidine rich intrinsically disordered proteins is regulated through zinc-histidine interactions," *Biomolecules*, vol. 9, no. 5, 2019.

[165] J. R. Long, W. J. Shaw, P. S. Stayton, and G. P. Drobny, "Structure and dynamics of hydrated statherin on hydroxyapatite as determined by solid-state NMR," *Biochemistry*, vol. 40, no. 51, pp. 15451–15455, 2001.

[166] G. Goobes, R. Goobes, O. Schueler-Furman, D. Baker, P. S. Stayton, and G. P. Drobny, "Folding of the c-terminal bacterial binding domain in statherin upon adsorption onto hydroxyapatite crystals," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 44, pp. 16083–16088, 2006.

[167] A. Amano, H. T. Sojar, J. Y. Lee, A. Sharma, M. J. Levine, and R. J. Genco, "Salivary receptors for recombinant fimbrillin of porphyromonas gingivalis," *Infect. Immun.*, vol. 62, no. 8, pp. 3372–3380, 1994.

Paper 1



Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

Carolina Cragnell, Ellen Rieloff and Marie Skepö

Division of Theoretical Chemistry, *Department of Chemistry, Lund University, P.O. Box 124, SE-221 00 Lund, Sweden*

Correspondence to Marie Skepö: *marie.skepo@teokem.lu.se*

https://doi.org/10.1016/j.jmb.2018.03.006
Edited by Jianhan Chen

Abstract

In this study, we have used the coarse-grained model developed for the intrinsically disordered saliva protein (IDP) Histatin 5, on an experimental selection of monomeric IDPs, and we show that the model is generally applicable when electrostatic interactions dominate the intra-molecular interactions. Experimental and theoretically calculated small-angle X-ray scattering data are presented in the form of Kratky plots, and discussions are made with respect to polymer theory and the self-avoiding walk model. Furthermore, the impact of electrostatic interactions is shown and related to estimations of the conformational ensembles obtained from computer simulations and "Flexible-meccano." Special attention is given to the form factor and how it is affected by the salt concentration, as well as the approximation of using the form factor obtained under physiological conditions to obtain the structure factor.

Introduction

Intrinsically disordered proteins and regions (IDPs and IDRs), from now on referred to as IDPs, are characterized by a lack of stable tertiary structure when the proteins exist as isolated polypeptide chains under physiological conditions *in vitro* [1,2]. More recently, it has been shown that ~30% of all proteins in eukaryotic organisms belong to this group of proteins, and that IDPs are involved in a large number of central biological processes and diseases. This discovery challenged the traditional protein structure paradigm, which stated that a specific well-defined structure was required for the correct function of a protein. Biochemical evidence has since shown that IDPs are functional, and that the lack of folded structures is related to their functions [3,4].

There is a great interest in the research community in the structure–function relationship for IDPs, and one hypothesis is that upon adsorption to surfaces, IDPs might adopt a structure, which gives rise to a function. Hence, for that purpose it is of interest to relate the properties of IDPs in solution with their properties in the adsorbed state, as well as their interaction with biological membranes. To be able to obtain a molecular understanding of macromolecules, it is useful to combine experimental techniques with atomistic and coarse-grained modeling. There have been great advances regarding atomistic simulations of IDPs, with the development and justification of force fields and water models, where the results have been validated against experimental results such as Förster resonance energy transfer, small-angle X-ray scattering (SAXS), and NMR. The reader is referred to the literature for more information [5–10]. The advantages of atomistic simulations are that one uses a full-atom approach and takes the water into account explicitly, whereas the limitation is that one is restricted to relatively short proteins due to the system size and computational power.

To be able to model longer proteins and more complex systems, coarse-grained modeling and Monte Carlo/molecular dynamics simulations are a good alternative. Of course, there will be approximations and simplifications; nevertheless, the approach has been shown to work very well. For more than 30 years, a coarse-grained model based on the primitive model [11], in combination with Monte Carlo simulations, has been used to model

Table 1. Details of the proteins within this study in terms of the length of the amino acid sequence, the number of phosphorylated residues (N_{phos}), the FCR, the NCPR, the percentage of prolines, and the number of hydrophobic residues (N_{hphob}). Furthermore, both the radii of gyration (R_g) obtained from experiments and simulations are included.

	Length	N_{phos}	FCR	NCPR	% Prolines	N_{hphob}	$R_{g,\ SAXS}$ (Å)	$R_{g,\ Sim}$ (Å)
Hst $5_{4–15}$ [16]	12	0	0.42	+0.42	0	2	9.2 ± 0.1	9.64 ± 0.02
Hst 5 [12]	24	0	0.38	+0.21	0	2	13.8 ± 0.1	13.77 ± 0.44
IB5 [15]	73	0	0.11	+0.08	40	5	27.9 ± 1.0	26.01 ± 0.05
Ash1 [13]	83	0	0.20	+0.18	15	12	28.4 ± 3.4	29.56 ± 0.02
Sic1 [14]	92	0	0.12	+0.12	16	20	28.8 ± 1.2	30.71 ± 0.05
II-1ng [15]	141	0	0. 19	+0.11	36	2	41.1 ± 1.0	38.24 ± 0.07
RNase E [17]	248	0	0.39	+0.05	6	55	52.6 ± 0.3	48.52 ± 0.11
Phosphorylated IDPs								
Statherin,	43	2	0.28	−0.09	16.3	7	19.3 ± 0.2	18.05 ± 0.05
pAsh1 [13]	83	10	0.45	−0.06	14.5	12	27.5 ± 1.2	21.76 ± 0.02
pSic1 [14]	92	6	0.25	−0.01	16.3	20	32.2 ± 2.2	27.55 ± 0.05

The experimental R_g values for Sic1 and pSic1 were determined using SAXS data obtained from the Protein Ensemble Database [14], and the Guinier approach.

polyelectrolytes and polyampholytes under various conditions. Sometimes this model is also referred to as the bead-necklace model. In this model, each monomer corresponds to a bead of a certain radius that can also have a charge associated with it. The water is always treated as a dielectric continuum.

In this study, we have used the coarse-grained model developed for the intrinsically disordered saliva protein Histatin 5 [12], on an experimental pool of IDPs obtained from different sources [13–18], as well as new experimental SAXS data for Statherin, also a saliva protein. We show that the model is generally applicable when electrostatic interactions dominate the intra-molecular interactions. For consistency, the reader should notice that we restrict our comparisons to experimental data obtained from SAXS. Focus will be on experimental and theoretically calculated SAXS data presented as Kratky plots, as well as comparison with polymer theory and the self-avoiding random walk (SARW) model. Furthermore, the impact of electrostatic interactions is shown and related to estimations of the conformational ensembles obtained from computer simulations and Flexible-meccano [19].

Results and Discussion

Polymer Model

The aim is to investigate if there exists a general coarse-grained model that accurately captures the structural properties of IDPs at both *high* and *low* salt concentrations. To assure the generality, the model developed for Histatin 5 [12] will be utilized on an experimental pool of IDPs covering a sequence length from 12 to 248 amino acids, and we will only compare the finding with experimental SAXS data. The IDPs have been characterized according to Das *et al.* [20], using the concepts: net charge per residue (NCPR), and fraction of charged residues (FCR), where NCPR = ($f_+ - f_-$) and FCR = ($f_+ + f_-$), with f being the fraction of positive/negative charges. According to this approach, polyampholytes and polyelectrolytes can be characterized to be either strong or weak, where FCR $\geq$ 0.3 corresponds to the former and FCR $<$ 0.3 to the latter. Moreover, they can be neutral, that is, NCPR $\approx$ 0, or have a net charge. Polyampholytes have approximately an equivalent fraction of opposite charges; thus, NCPR is low, whereas polyelectrolytes have more of one type of charge. The proteins used in this study are summarized in Table 1. As shown, although the selection of proteins might seem small, a fairly representative pool of IDPs is given with respect to the charges, the number of phosphorylated residues (N_{phos}), the number of hydrophobic amino acids (N_{hphob}), and the proline content. The number of hydrophobic residues is based on the notion that all amino acids with a higher hydropathy value than glycine in the Kyte–Doolittle scale [21], are considered hydrophobic.

The level of compaction/extension has been analyzed by comparing the radius of gyration (R_g) from SAXS with the corresponding analysis obtained from Monte Carlo simulations, that is, comparison of ensemble-averaged estimates as well as the full conformational ensemble through the probability distribution. Fig. 1a displays the radii of gyration from the simulations *versus* the experimental counterparts. As is clearly shown, there is a good correspondence between the ensemble estimates. However, there are proteins that display simulated radii of gyration that are statistically different from the experimental data; moreover, the experimental data are more extended than the model predicts, that is: RNase E, two of the phosphorylated proteins, namely, pAsh1, and pSic1, as well as the proline-rich protein II-1ng. For RNase E, we hypothesize that it is due to a slight degree of self-association; for pAsh1 and pSic1, we expect it to be

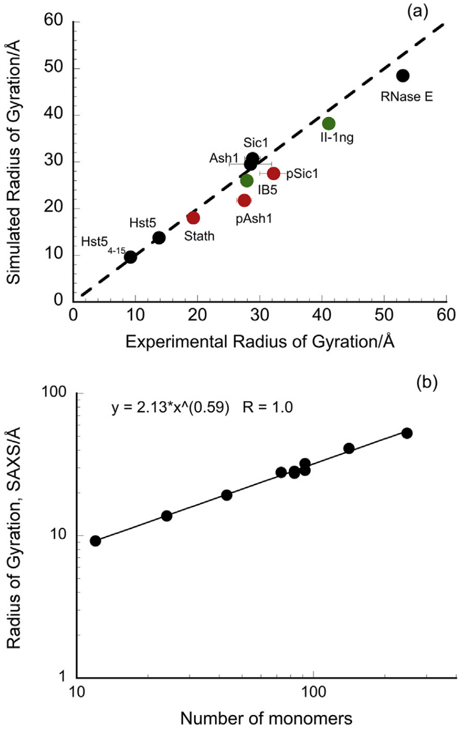


Fig. 1. (a) Radii of gyration obtained from simulations *versus* the radii of gyration obtained from experiments where black filled circles correspond to non-phosphorylated IDPs, red filled circles to phosphorylated proteins where the phoshate group is assumed to have a net charge of $-2e$, and green filled circles to proline-rich proteins. (b) The experimental radii of gyration as a function of the protein sequence length on log–log scale. The ionic strength corresponds to 150 mM, except for IB5 and II-1ng, where it was 50 mM. For most of the reported values, the precision is smaller than the marker in the plot; hence, the reader is referred to Table 1 for more information.

due to the high number of phosphorylated residues, whereas for II-1ng it is due to the proline content which, due to the cyclic structure of the amino acid, gives the proline an exceptional conformational rigidity. Nevertheless, the reader should notice that the radii of gyration for the proline-rich proteins do agree remarkably well.

For some polymers, such as the well-known polymer polyethylene glycol, it is possible to define an empirical expression for a simplistic estimation of the R_g [22], according to the power-law $R_g = \rho_0 N^{\upsilon}$. In this context, υ refers to the Flory exponent, which depends on the structural behavior of the polymer chain in the solvent, N refers to the number of monomers in the chain, and ρ_0 is a prefactor. The latter is a function of, among other things, the details of the monomer as the radius, the persistence length, and the bond geometry. This leads to the question: Is it possible to define a similar expression for IDPs as for polyethylene glycol? For a random walk (also denoted ideal chain), the parameter υ is equal to 0.5, whereas it is approximately 0.6 for a SARW [23]. In the latter, the interactions between the chain monomers (or for IDPs, the amino acids), are modeled as excluded volumes, which cause a reduction in the conformational possibilities of the chain, in comparison with a random walk where all bonds and torsion angles are equally probable. In Fig. 1b, the experimentally obtained radii of gyration (from SAXS) of our selection of model proteins are shown as a function of sequence length. From the fit to the curve, υ is estimated to be approximately 0.59, which matches closely the exponent obtained from the computer simulations ($\upsilon = 0.58$), where only excluded volumes are taken into account (data not shown). Hence, it seems that the selection of IDPs used in this study behave as SARWs under the given solution conditions, that is, high ionic strength. This is a reasonable conclusion when electrostatic interactions dominate the intra-chain interactions, which can be highly screened by the large amount of salt present in the solution. This rationale is further verified since the fractions of hydrophobic residues of the used IDPs are rather low, $\leq 20\%$ (see Table 1).

By fitting the experimentally obtained radii of gyration as a function of the number of amino acids for the proteins used in this study, we obtain a prefactor ρ_0 of approximately 2.13, which is in good agreement with the model in the computer simulations where the radius of the amino acids is set to 2 Å. In the literature, the Flory exponent varies between $\upsilon = 0.5$ and 0.6 depending on the technique (Förster resonance energy transfer or SAXS), protein, and solvent used, that is, in the latter with or without denaturing agents [24–30]. This is plausible since the Flory exponent is sensitive to the intramolecular interactions in the protein, thus the amino acid composition. A more hydrophilic protein with a low fraction of hydrophobic amino acids will obtain the higher value of the Flory exponent, whereas the opposite occurs if the fraction of charges is low and the number of hydrophobic amino acids is high, where the latter has been reported by Hofmann *et al*. [27]. It is very interesting to notice though that hydrophobic disordered proteins are expanded in water, as reported for example by Riback *et al*. [31]. In the latter, the authors of this paper hypothesize that the decrease in the Flory exponent might be due to the hydrophobic effect; that is, the final conformational state is driven by the total minimization of the hydrophobic surface, which manifests itself as an effective attractive force. Notice also that the statistical basis in all experimental studies presented is rather low; hence, the shape of the curve is rather sensitive to the addition of a further IDP.

As is well known, an IDP can exist in an infinite number of spatial states due to its high flexibility and fast dynamics. To obtain more information about the conformational averages, the Monte Carlo simulation technique is invaluable since it gives the Boltzmann-weighted probability of finding a system in a specific state. The properties of IDPs are of course dependent on different parameters such as the amino acid sequence and the temperature, as well as the solution properties. It has been shown in several papers [27,28,30], and above, that the IDPs can be considered to behave as SARWs when only steric interactions are taken into account due to high salt concentration or the presence of a denaturing agent. The next question is: How does the chain length affect the conformational ensemble average under such conditions? For this purpose, we have analyzed the full width half maximum (FWHM) and peak position of the probability distribution function of the radius of gyration and the shape of the adopted conformations using our model protein without charges, that is, considering only steric interactions. As expected and shown in Fig. 2, the ensemble of possible conformations increases as a function of the number of amino acids; cf. R_g spans from 10 to 35 Å, and from 40 to 130 Å, for 50- and 500-amino-acid monomers, respectively. By analyzing the FWHM as a function of the number of amino acids in the protein sequence, an estimate of the conformational entropy of the model protein can be obtained such that the broader the peak, the larger the chain entropy. The FWHM and the peak position as a function of protein length show the same $\upsilon \approx 0.6$ scaling behavior as the radius of gyration (data not shown).

The shape of the IDP can be defined as the ratio of the mean-square end-to-end distance, $\langle R_{ee}^2\rangle^{1/2}$, and the mean-square radius of gyration $\langle R_g^2\rangle^{1/2}$ (also denoted R_{ee} and R_g) according to: $r_{shape} = \langle R_{ee}^2\rangle / \langle R_g^2\rangle$. In the rod-like limit, $r_{shape} = 12$; for a flexible chain in good solvent, $r_{shape} \approx 6.3$; and for an ideal chain, $r_{shape} = 6$. For all chain lengths, the shape probability distribution is a symmetric bell-shaped function with a broad maximum of only 0.15 at $r_{shape} = 6$. The latter number indicates that a specific average conformation occurs during 15% of the simulation length (data not shown). Hence, there is a relatively high probability to accommodate all the different possible shapes, for example, from a rather contracted chain to a rigid prolate. Notice that $r_{shape} = 1$ does not necessarily indicate that an IDP is a compact globule, rather that the chain is contracted and that the mean-square end-to-end distance and the mean-square radius of gyration are of the same order.

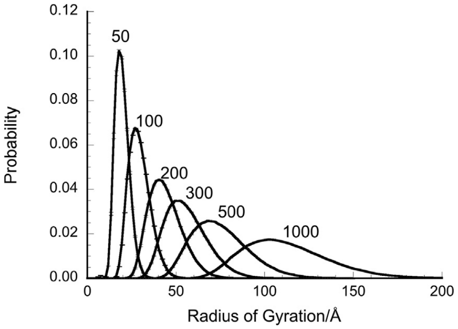


Fig. 2. The conformational ensemble of radius of gyration for different lengths of the model protein, where only steric interactions through excluded volumes are taken into account.

The effect of electrostatic interactions on the single molecular level

The impact of electrostatic interactions at the single molecular level on the conformational ensemble of IDPs, and how it affects the scattering spectra, visualized as Kratky plots, has also been investigated. Of particular interest is when the ionic strength is 150 mM, since that is commonly applied in SAXS experiments to determine the form factor. Here, the study has been divided into two parts: (i) non-phosphorylated and (ii) phosphorylated proteins.

Non-phosphorylated IDPs

Fig. 3 shows the obtained radii of gyration calculated from simulations at 10 mM and 150 mM salt, which corresponds to Debye screening lengths

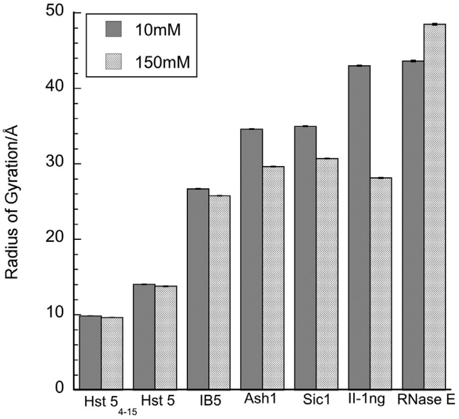


Fig. 3. The simulated radii of gyration of the chosen IDPs at high and low ionic strength (150 and 10 mM).

(κ^{-1}) of approximately 30 and 8 Å, respectively. As shown, it is clearly visible that upon the addition of salt, some proteins attain polyelectrolytic behavior, whereas other proteins exhibit polyampholytic behavior. In the former, the protein contracts, whereas in the latter, it becomes more extended when the salt concentration is increasing. Moreover, a clear trend is also obtained with respect to the chain length; that is, the screening effect is more accentuated for longer proteins, which induces larger discrepancies in the estimated extensions. Hence, in this respect, the charge distribution obtained from the specific amino acid sequence and the protein length due to the higher probability to attain a larger population of conformations are of importance.

The effect of salt on R_g and the conformational ensemble has been further analyzed focusing on the protein Ash1$_{420–500}$ (hereafter referred to as Ash1). This protein has been extensively studied in the paper by Martin *et al*. [13]. Among other things, they showed that Ash1 adopts coil-like conformations that are expanded and well solvated. The R_g for Ash1 from experiments and modeling with and without charges at different ionic strengths are given in Table 2. There is a clear trend in the simulated R_g, which decreases as a function of salt concentration. The SAXS measurements (150 mM salt) gave an R_g of 28.5 ± 3.4 Å, which means that all simulated radii of gyration except the one obtained at 10 mM salt are within the uncertainty. The simulations show that the conformational properties of SARW are reached first upon the addition of 1000 mM salt, that is, when the Debye screening length is shorter than the average bead-to-bead distance in the model, cf. 3.04 Å for the former with 4.1 Å for the latter. The reader should notice that the more dramatic effects occur, of course, in the lower salt regime, for example, between 10 and 150 mM salt. These results are clearly shown in the probability distribution of the conformational ensemble as given in Fig. 4a. Notice that a small change in the ensemble average will affect the conformational ensemble more remarkably, and that the electrostatic interactions within the chain are quite pronounced even at higher salt concentrations. As shown in Fig. 4a, Ash1 behaves as a polyelectrolyte in the sense that it contracts upon the addition of salt. The FWHMs of the probability distribution of R_g for Ash1 at an ionic strength of 10 and 150 mM are estimated to be 13.70 ± 0.10 and 12.91 ± 0.18 Å, respectively. These numbers confirm that the conformational entropy of Ash1 is decreasing upon salt addition, which is in line with the fact that the preferred shape is more contracted at higher salt concentrations.

Table 2. Conformational properties and the FWHM of the IDR in Ash1 as a function of salt obtained from simulation.

I (mM)	κ^{-1} (Å)	R_g (Å)	R_{ee} (Å)	FWHM (Å)
10	30.4	34.54 ± 0.01	88.43 ± 0.05	13.70 ± 0.10
150	7.9	29.56 ± 0.02	74.33 ± 0.05	12.91 ± 0.18
300	5.6	28.68 ± 0.02	71.99 ± 0.05	12.71 ± 0.19
500	4.3	28.19 ± 0.02	70.69 ± 0.06	12.63 ± 0.20
1000	3.04	27.77 ± 0.01	69.62 ± 0.04	12.58 ± 0.20
SARW	N.A.	27.28 ± 0.04	68.12 ± 0.13	12.47 ± 0.21
SAXS	7.9	28.5 ± 3.4	N.A.	N.A.

Included also is the radius of gyration obtained from SAXS by Martin *et al*. [13] at an ionic strength of 150 mM and the simulated SARW for Ash1.

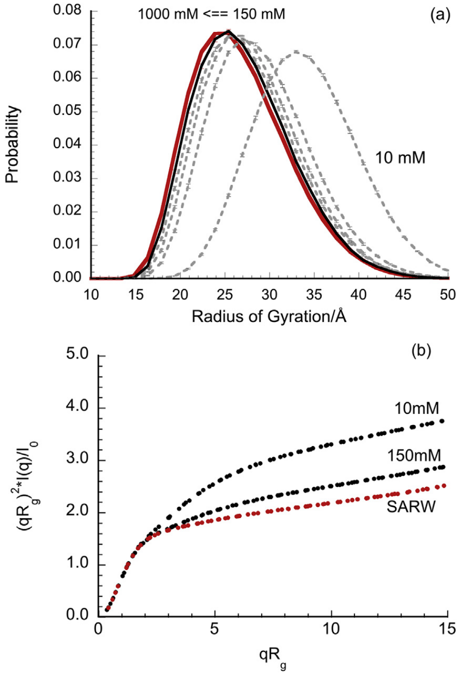


Fig. 4. The probability distribution of the radius of gyration (a), that is, conformational ensemble, and the dimensionless Kratky plot as a function of salt concentration for Ash1 (b). The red function corresponds to the SARW, whereas 10 and 150 mM are shown as black-dotted curves. In panel a, the full black line corresponds to 1000 mM.

The asphericity ranges from 0 for a sphere to 1 for a rod, and have been determined according to the protocol by Angelescu and Linse [32]. The ensemble averages of the asphericity as well as the shape factor indicate that at low ionic strength, that is, 10 mM, Ash1 becomes more extended than a SARW, the values being 0.6 and 6.6, respectively. At increased salt concentrations, the values level off to approximately 0.5 for the asphericity and 6.3 for the shape, clearly indicating conformations resembling a SARW. Hence, at 150 mM and higher ionic strengths, it is possible to model the form factor as a SARW, especially when taking into account the resolution of SAXS experiments. However, it is important to remember that it is indeed an approximation, as true SARW behavior is reached first at 1000 mM. At low salt concentration, it is not possible to model the form factor as a SARW, and additionally, the differences between 10 and 150 mM are quite pronounced. On the other hand, it is also very difficult to measure the form factor of IDPs at low salt concentrations by SAXS due to the contribution from the structure factor on the scattering curve. An advantage with computer simulations is that it enables discrimination of how intra- and intermolecular interactions affect the form factor. Fig. 4b shows the unitless Kratky plot that qualitatively assesses the overall conformational state and reveals the flexibility/rigidity of the protein. Both the results obtained from simulations at 10 and 150 mM salt, as well as for a SARW, are shown for comparison. In this representation, the salt effect is clearly visible and these results confirm, indeed, that the form factor depends on the salt concentration; that is, it is not accurate to use the same form factor at *high* and *low* ionic strength. This will of course have implications when deriving the structure factor at low ionic strengths using: $I(q) = S(q) \cdot P(q)$, where $P(q)$ often is determined at a higher salt concentration by SAXS. Here $S(q)$ and $P(q)$ correspond to the structure and the form factor, respectively.

Phosphorylated IDPs

Many of the IDPs belong to the family of phosphoproteins; that is, for example, they often contain phosphorylated serines or threonines. In this study, three model proteins have been investigated: Statherin, pSic1, and pAsh1. The first protein, Statherin, contains two phosphorylated serines residing in the N-terminus, possesses an amphiphilic structure, and has a tendency to self-associate. In the second protein, pSic1, there are six phosphorylated groups, whereas in pAsh1, there are ten. The reader is referred to Fig. 5 to achieve an overview of the distribution of the phosphorylated as well as the positively and the negatively charged amino acids. Furthermore, according to Das *et al.* [20], FCR and NCPR (denoted FCR:NCPR) for Statherin, pSic1, and pAsh1 are 0.23:−0.05; 0.25:−0.01; and 0.46:−0.06. Hence, the two former can be considered as weakly charged polyelectrolytes/polyampholytes where pSic1 is almost net neutral, whereas in this context, pAsh1 is strongly charged. As a reminder, the threshold for strongly charged polyelectrolytes is FCR > 0.3.

Starting off with Statherin, our SAXS measurements show that despite its tendency to self-associate, it is possible to obtain a form factor for Statherin at low protein concentrations. As shown in Fig. 1a as well as given in Table 1, the experimentally and simulated radii of gyration agree relatively well; hence, the two phosphorylated serines at position 2 and 3 do not seem to influence the ensemble average to greater extent in that respect. Fig. 6a shows the dimensionless Kratky plot, and as clearly visible, the profiles from the experiment and the simulation agree very well and display a random coil behavior, that is, a linear rise to a plateau at higher scattering angles. Interestingly, the simulation snapshots indicate that the N-terminus where the two phosphorylated serines reside seems to form a cluster, while the rest of the chain is flexible, as illustrated by Fig. 6b. From the simulations, it is also shown that the R_g is not sensitive to salt (data not shown).

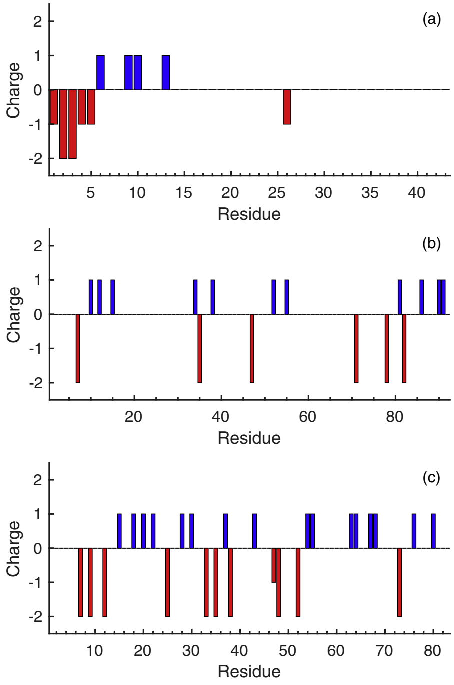


Fig. 5. Charge distribution at pH 7 for Statherin (a), pSic1 (b), and pAsh1 (c), where positive charges are marked in blue and negative charges in red. The N- and C-terminal charges are not included.

pSic1 on the other hand is twice as long as Statherin and contains six phosphorylated residues at positions 7, 35, 47, 71, 78, and 82, that is, relatively well separated from each other. As shown in Fig. 1a, there is a significant difference in the radii of gyration obtained from the experiment *versus* the simulation, where the former indicates a conformation more expanded than a SARW, and the latter displays a more compact conformation, less expanded than SARW (28.94 ± 0.05 Å). From the simulations, it is

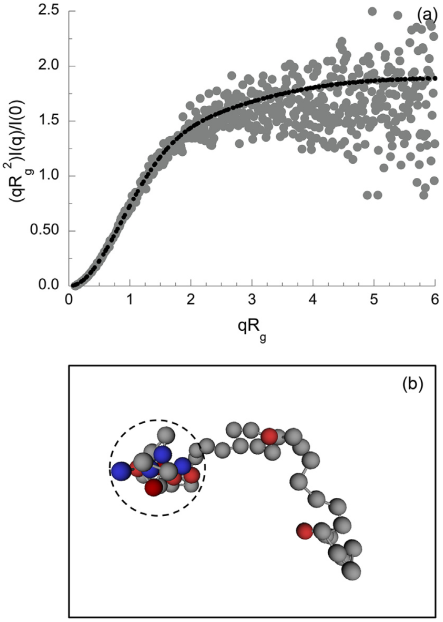


Fig. 6. (a) Dimensionless Kratky plot for experimental data at pH 8.1 (gray filled circles) and for the simulated data (black filled circles) at an ionic strength of 150 mM for Statherin. (b) Representative snapshot of a chain conformation obtained in a simulation at 150 mM salt. Blue spheres are positively charged amino acids, red spheres are negatively charged amino acids, and the dark red spheres represent phosphorylated serines with the charge $Z_{\mathrm{phos}} = -2e$, whereas the gray spheres correspond to neutral amino acids. The salt was treated implicitly, and the counterions are omitted for clarity. The dashed line circles the N-terminal part of the chain.

also shown that R_g is sensitive to salt and decreases when the salt concentration is increased, from 31.11 ± 0.05 Å to 27.55 ± 0.05 Å at 10 and 150 mM salt, respectively, which advocates the existence of electrostatic attractive interactions within the chain.

The last phosphorylated protein in our study, pAsh1, contains 10 phosphorylated residues, where nine out of ten are within the 52 amino acids in the N-terminal (positions 7, 9, 12, 25, 33, 35, 38, 48, 52, and 74). As shown in Fig. 1a, there is a discrepancy between the experimental and simulated data, where the simulation again advocates a more contracted ensemble average than the experiment as well as SARW (27.28 ± 0.04 Å). Experimentally, it has been shown that upon phosphorylation of Ash1 at ten distinct sites, the global conformational properties of pAsh1 are indistinguishable from those of unphosphorylated Ash1. The obtained ensemble averages of the radii of gyration from SAXS measurements were determined to be 28.4 ± 3.4 Å and 27.5 ± 1.2 for Ash1 and pAsh1, respectively, at 150 mM NaCl [13]. Simulations of the ensemble average of the radius of gyration as a function of salt clearly indicate that Ash1 displays a polyelectrolytic and pAsh1 a polyampholytic behavior (see Fig. 7) and that realistic trends are captured.

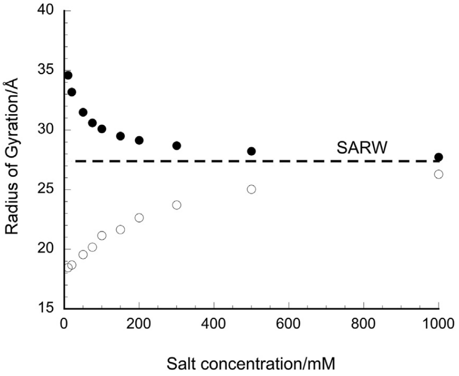


Fig. 7. The ensemble average of the radius of gyration in Å as a function of the salt concentration in mM, for Ash1 in black circles and the 10-sites phosphorylated counterpart pAsh1 in open circles. The salt is assumed to be of 1:1 nature with respect to the charge. The dashed line corresponds to the estimated radius of gyration utilizing the SARW. The reader should keep in mind that the experimentally obtained values of R_g for the two proteins correspond to 28.5 ± 3.4 Å and 27.5 ± 1.2 Å, [13], respectively, which is approximately the same number as obtained from the SARW model. The precision of the data is too small in comparison with the marker to be visible.

Our conclusion is that depending on the number of phosphorylated sites and their distribution, short-ranged attractive electrostatic interactions could influence the conformational properties quite dramatically. For Ash1/pAsh1, the radius of gyration decreases with ≈ 26%, whereas the corresponding numbers for Sic1/pSic1 and Statherin system are 10% and 1%, respectively. Moreover, the shape of the proteins deviates more dramatically when phosphorylated groups are introduced, cf. protein with and without phosphorylation. The effect is enhanced with an increasing number of phosphorylated residues, as visualized in the Kratky plots obtained from simulations in Fig. 8. The dependence of the amino acid distribution is further strengthened by the partial radial distribution function between the positively charged amino acids and the phosphorylated residues in Fig. 9, which emphasizes the effect of short-ranged attractive electrostatic interactions. Moreover, as shown in Fig. 10, a substantial amount of salt is needed to screen this short-ranged attractive electrostatic interaction; that is, κ^{-1} needs to be shorter than the distance between the amino acids within the chain.

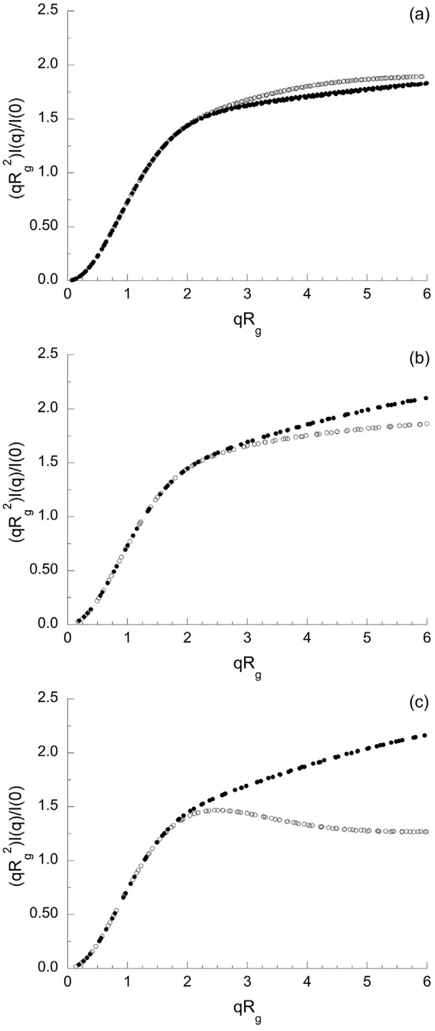


Fig. 8. The simulated dimensionless Kratky plot for Statherin with and without phosphorylated residues (a), Sic1/pSic1 (b), and Ash1/pAsh1 (c), where open circles represent the phosphorylated protein and filled circles the non-phosphorylated counterpart. The reader should notice that the number of phosphorylated groups is increasing from two to six to ten, for the phosphorylated proteins in panels a, b, and c, respectively.

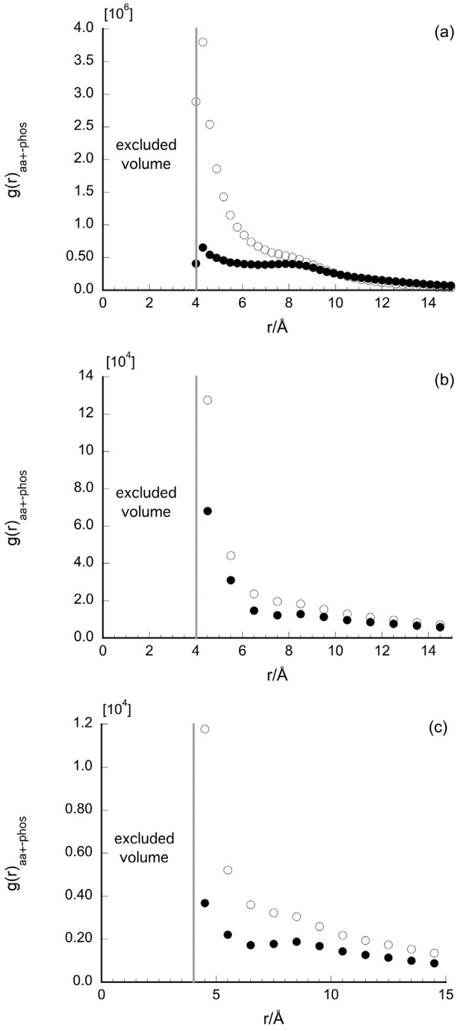


Fig. 9. Partial radial distribution function between positively charged amino acids and phosphorylated residues at 150 mM salt for Statherin (a), pSic1 (b) and pAsh1 (c), where the phosphate groups have the charge $-2e$ (open circles) or 0 (filled circles, corresponding to non-phosphorylated protein).

A plausible explanation to the difference between the experimental and simulated radius of gyration for pAsh1 could be due to the physicochemical properties of the phosphorylated residue. Phosphorylation changes the characteristics of the amino acids, especially due to introducing charge. The first pK_a of the phosphate group is below 3, while the second pK_a value is slightly below 6 [33,34], meaning that at physiological pH, the phosphate group should carry a $-2e$ charge. However, pK_a values between 6.9 and 7.2 have also been found in Web-based tools for calculating the point of zero charge (see http://scansite.mit.edu/calc_mw_pi.html and ProMoST) [35]. Hence, the radius of gyration has also been determined by simulating the corresponding proteins

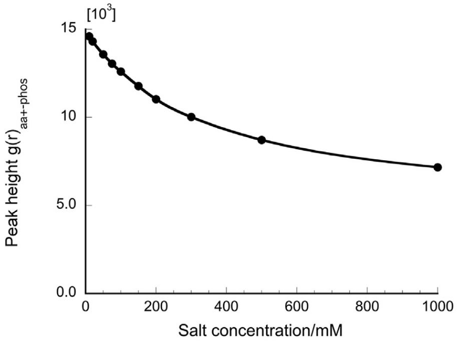


Fig. 10. Peak value of the partial radial distribution function at 4.5 Å between positively charged amino acids and phosphorylated residues as a function of salt concentration, for pAsh1. The precision is within the data marker.

for the phosphorylated proteins where the phosphate group carries a charge $Z_{\text{phos}} = -1e$. As shown in Table 3, it gives a much better agreement with the experiments. However, no such interpretation should be made as the phosphorylated residues carry the charge $Z_{\text{phos}} = -1e$ at physiological pH. Other possibilities could be that there is a distribution of phosphorylated residues in the experimental sample which does not exist in the model, or that some phosphorylated residues are neutralized due to their binding affinity to, for example, calcium. Monte Carlo simulations provide an exact solution to the model used; hence, traces of other proteins, multivalent ions, and so on, do not exist, which should be kept in mind when comparison are performed with the experimental counterpart.

Model adjustability

The total potential energy of the coarse-grained model presented in this study includes a short-ranged attractive interaction between all amino acids, as well as explicit charges depending on the nature of the amino acid. Moreover, the protein is modeled as totally flexible in the sense that steric interactions are included only through the excluded volume of the amino acid; that is, the chain entropy might be overestimated and the protein too fluidic. This can, of course, be opposed by introducing, for example, an angular potential or increasing the amino acid excluded volume to decrease the flexibility, which is of relevance for the group of proline-rich proteins. Here we compare our modeling results with the non-glycosylated proline-rich saliva proteins, IB5 and II-1ng [15], whose amino acid sequences contain approximately 40% prolines. The experimental and simulated radii of gyration are approximately equivalent, taking the uncertainties into consideration. Although the radius of gyration agrees very well, that might not be the case for the shape. This will be further analyzed by focusing on IB5. As shown in the Kratky plot in Fig. 11, there is a discrepancy between the experimental and simulated curves. From the experimental Kratky profile, one can conclude that the ensemble is biased toward more stiff conformations, in comparison to the unperturbed model (black curve), which, most probably, is an effect of the high proline content.

One possibility to improve the agreement between SAXS and simulations is by introducing an angular potential. The effect of the prolines has been taken into account in the simulations by adding an angular potential of 0.0023 kJ mol^{-1} deg^{-2}; that is, the average angle between three consecutive beads increased from approximately 103° to 141°, that is, a quite dramatic change (see red curve). The resulting radius will then be overestimated but the flexibility/rigidity is more realistic. Another possibility would be to induce a local stiffness within the chain representing

Table 3. Number of phosphorylated residues, N_{phos}, and simulated radii of gyration (R_g) for phosphorylated IDPs, expressed in Å, at 150 mM monovalent salt for phosphorylated residues with the net charge of $Z_{\text{phos}} = -1e$ or $Z_{\text{phos}} = -2e$

	N_{phos}	$R_{g,\text{ exp}}$ [Å]	$R_{g,\text{ sim}}$ [Å] $Z_{\text{phos}} = -1$	$R_{g,\text{ sim}}$ [Å] $Z_{\text{phos}} = -2$
Statherin	2	19.3 ± 0.2	18.24 ± 0.04	18.05 ± 0.05
pSic1	6	28.6 ± 0.5	29.00 ± 0.06	27.55 ± 0.05
pAsh1	10	27.5 ± 1.2	25.61 ± 0.08	21.66 ± 0.12

The experimental SAXS data for pAsh1 and pSic1, respectively, are obtained from Martin *et al.* [13] and Mittag *et al.* [39].

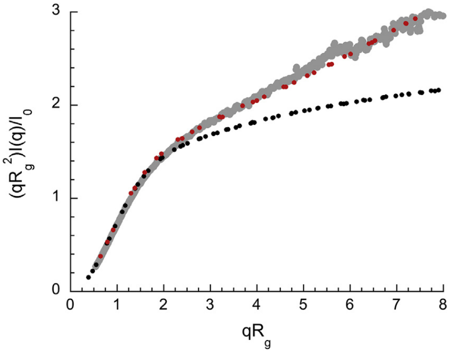


Fig. 11. Dimensionless Kratky representation of IB5 from SAXS measured by Boze *et al.* [15] (gray), the flexible protein model (black), and the model with an additional angular potential, k_{angle} = 0.0023 kJ mol^{-1} deg^{-2} (red).

the segments consisting of several prolines. This is, however, out of the scope for the current study, since we are aiming for a general model, which can be easily adjusted to all IDPs with a few parameters.

Conclusions

To summarize our findings, the coarse-grained model, based on the primitive model, is well applicable for IDPs where the intra-chain interactions are dominated by electrostatic interactions. By extending the model to include, for example, angular potentials, and/or a short-ranged attractive interaction preferably between the hydrophobic amino acids within the chain, in principle it is possible to tune the fitting parameters to obtain an agreement between the simulations and the experimental data for a specific protein.

A popular method for analyzing SAXS spectra of IDPs and to achieve information about the ensemble average of the radius of gyration is by utilizing Flexible-meccano. Comparisons between the results obtained from Monte Carlo simulations and Flexible-meccano agree well. As shown in Fig. 12, this method works well for the unphosphorylated IDPs used in this study and it is definitely a valuable tool to obtain information about the most probable conformations and R_g distributions. The take-home message is that coarse-grained modeling and Monte Carlo simulations can contribute when the aim is to understand the underlying physics and the intricate balance between the different contributions regarding the intra-chain interactions. The model seems to be generally valid when electrostatic interactions dominate, and it can be adjusted to correspond to any IDP/IDR by tuning the intra-chain potentials.

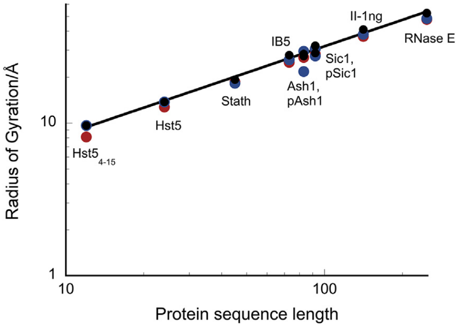


Fig. 12. The ensemble average of radius of gyration as a function of the length of the amino acid sequence in the protein on a log–log scale for the experimental pool of proteins where the full line including black data markers corresponds to a power law fit of the experimental values, the red filled circles to the results obtained from Flexible-meccano, and the blue filled circles from Monte Carlo simulations.

Furthermore, it is possible to use an empirical expression to achieve an estimate of the radius of gyration of the monomeric protein when the dominant intra-chain interactions are electrostatic in nature. This could be of practical importance when performing experiments to achieve a rapid understanding of, for example, the association state of the protein or if there exist residual elements of local structure.

Coarse-grained modeling and Monte Carlo/molecular dynamics simulations are valuable approaches when the aim is to achieve an understanding of how the structure and the inter- and intramolecular interactions are affected by variations in pH, salt concentration, and protein point mutations. It is also useful for studying more complex systems, such as the effect of protein concentration, interaction with other macromolecules (e.g., proteins and surfactants), as well as the interaction with surfaces and biological membranes. In the latter, the distribution and valency of the surface charges, the surface charge density, and the bilayer composition can be evaluated. The information from these simulations can then be correlated with the function.

Model and Method

Coarse-grained model

The monomers of the proteins, that is, the amino acids, are represented by hard spheres (beads) that mimic their excluded volume including the hydration layer and are connected via harmonic bonds. The N- and C-termini are included explicitly to account for the extra charge. The bead radius was set to 2 Å providing a realistic contact separation between the charges and an accurate Coulomb interaction. The non-bonded spheres interact through a short-ranged attractive interaction and electrostatic interactions, where the interparticle electrostatic interactions are described on the Debye–Hückel level. The simulations are performed at constant pH with point charges. Each monomer is negative, positive, or neutral, depending on the amino acid sequence, as illustrated in Fig. 13.

The total potential energy of the simulated system contains bonded and non-bonded contributions, and is given by:

$$U_{\mathrm{tot}} = U_{\mathrm{nonbond}} + U_{\mathrm{bond}} = U_{\mathrm{hs}} + U_{\mathrm{el}} + U_{\mathrm{short}} + U_{\mathrm{bond}} \tag{1}$$

where the non-bonded energy is assumed to be pairwise additive according to:

$$U_{\mathrm{nonbond}} = \sum_{i<j} u_{ij}(r_{ij}), \tag{2}$$

where $r_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ is the center-to-center distance between two monomers, and $\mathbf{R}$ refers to the

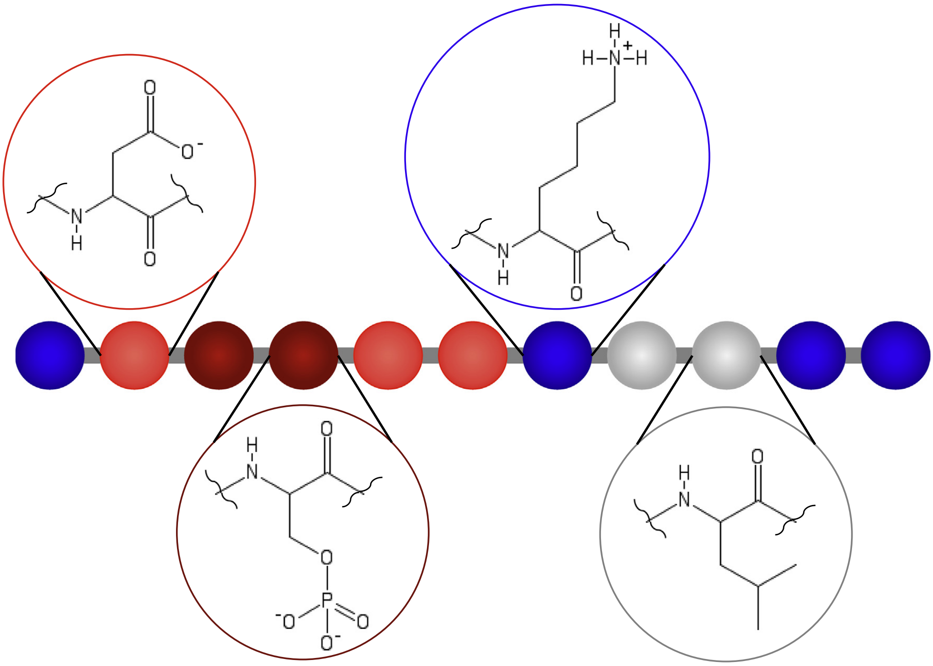


Fig. 13. Schematic description of the coarse-grained model showing the N-terminal fragment of the saliva protein Statherin. Blue spheres have the charge $Z = +1e$; bright red spheres, $Z = -1e$; and dark red spheres, $Z = -2e$. Gray spheres correspond to neutral amino acids. The four structures depicted are aspartic acid, phosphorylated serine, lysine, and leucine. The N-terminal is modeled explicitly as a positively charged sphere.

coordinate vector. The excluded volume is taken into account through the hard-sphere potential, U_{hs}, given by:

$$U_{\mathrm{hs}} = \sum_{i<j} u_{ij}^{\mathrm{hs}}(r_{ij}), \tag{3}$$

which sums up over all amino acids. The hard-sphere potential, $u_{ij}^{\mathrm{hs}}(r_{ij})$, between two monomers in the model is given by:

$$u_{ij}^{\mathrm{hs}}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j \\ \infty, & r_{ij} < R_i + R_j \end{cases}, \tag{4}$$

where R_i and R_j denote the radii of the beads. The electrostatic potential U_{el}, is given by an extended Debye–Hückel potential according to:

$$U_{\mathrm{el}} = \sum_{ij} u_{ij}^{\mathrm{el}}(r_{ij}) = \sum_{i<j} \frac{Z_i Z_j e^2}{4\varepsilon_0 \varepsilon_r} \frac{\exp[-\kappa(r_{ij} - (R_i - R_j))]}{(1+\kappa R_i)(1+\kappa R_j)} \frac{1}{r_{ij}}, \tag{5}$$

where e is the elementary charge, κ denotes the inverse Debye screening length, ε_0 is the vacuum permittivity, and ε_r the dielectric constant for water. The short-ranged attractive interaction between the monomers is included through an approximate arithmetic average over all amino acids, given by:

$$U_{\mathrm{short}} = -\sum_{i<j} \frac{\varepsilon}{r_{ij}^6}, \tag{6}$$

where ε reflects the polarizability of the proteins and thus sets the strength of the interaction. In this model, ε was set to 0.6 × 10^4 kJ Å^6/mol giving an attractive potential of 0.6 kT at closest contact. The bonded interaction, a harmonic bond, is given by:

$$U_{\mathrm{bond}} = \sum_{i=1}^{N-1} \frac{k_{\mathrm{bond}}}{2} (r_{i,i+1} - r_0)^2 \tag{7}$$

where $\boldsymbol{r}_{i,i+1}$ denotes the distance between two connected monomers with the equilibrium separation $r_0 = 4.1$ Å, and the force constant $k_{\mathrm{bond}} = 0.4$ N/m, whereas N denotes the number of monomers of the protein. The proteins are assumed to be totally flexible, except for when the effect of intrinsic stiffness is evaluated. An angular dependent component, expressed below, is then added to the potential:

$$U_{\text{angle}} = \sum_{i=2}^{N-1} \frac{k_{\text{angle}}}{2} (\alpha_i - \alpha_0)^2. \tag{8}$$

Here, α_i is the angle formed by the vectors $\boldsymbol{r}_{i+1} - \boldsymbol{r}_i$ and $\boldsymbol{r}_{i-1} - \boldsymbol{r}_i$, made by three consecutive beads with the equilibrium angle $\alpha_0 = 180°$ and the force constant k_{angle}. In addition to the angular potential, the electrostatic interactions among the segments as well as the volume of the hard spheres also contribute to the rigidity of the protein.

Simulation aspects

The equilibrium properties of the model systems were obtained applying Monte Carlo simulations in the canonical (NVT) ensemble, that is, constant volume, number of beads, and temperature ($T = 298$ K), utilizing the Metropolis algorithm. The protein chain was enclosed in a cubic box of variable volume, which was dependent on the protein length. Periodic boundary conditions were applied in all directions. The long-ranged Coulomb interactions were truncated using the minimum image convention. Four different types of displacements were allowed: (i) translational displacement of a single bead, (ii) pivot rotation, (iii) translation of the entire chain, and (iv) slithering move, in order to accelerate the examination of the configurational space [36]. The probability of the different trial moves was weighted to enable single-particle moves 20 times more often than the other three. Initially, the protein was randomly placed in the box and an equilibrium simulation of typical 2×10^5 trial moves/bead was performed, whereas the proceeding production run comprised 10^6 passes divided into 10 subdivisions. The radius of gyration and end-to-end distance probability distribution functions of the proteins, that is, the conformational ensembles, were analyzed to confirm that the simulations were sampled accurately. The reported uncertainty of simulated quantities is one standard deviation of the mean. It is estimated from the deviation among the means of the subdivisions of the total number of MC passes according to:

$$\sigma^2(\langle x \rangle) = \frac{1}{n_s(n_s - 1)} \sum_{s=1}^{n_s} \left(\langle x \rangle_s - \langle x \rangle\right)^2, \tag{9}$$

where $\langle x \rangle_s$ is the average of quantity x from one subdivision, $\langle x \rangle$ the average of x from the total simulation, and n_s the number of subdivisions. The simulations were performed by using the integrated Monte Carlo/molecular dynamics/Brownian dynamics simulation package Molsim [37].

Structural analysis

The model was validated by comparing the simulated scattering intensities with the experimental scattering intensities obtained by SAXS. For a system containing N identical scattering objects, the structure factor is given by:

$$S(q) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp\left(i\mathbf{q} \cdot \mathbf{r}_j\right) \right|^2 \right\rangle. \tag{10}$$

The total structure factor can further be decomposed into partial structure factors given by:

$$S_{\text{ij}}(q) = \left\langle \frac{1}{\left(N_i N_j\right)^{1/2}} \left[\sum_{i=1}^{N_i} \exp(i\mathbf{q} \cdot \mathbf{r}_i)\right] \left[\sum_{j=1}^{N_j} \exp(-i\mathbf{q} \cdot \mathbf{r}_j)\right] \right\rangle. \tag{11}$$

The total and partial S(q) are related through:

$$S(q) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \frac{\left(N_i N_j\right)^{1/2}}{N} S_{ij}(q). \tag{12}$$

For a point scatterer, the form factor is constant, inferring that the scattering intensity is proportional to the structure factor. In order to account for an approximate effective particle/residue form factor, the scattering profile further needs an appropriate normalization, such that I_0 coincides with the experimental scattering profile.

FWHM analysis

To obtain the FWHM of the radius of gyration probability distribution, the curve was fitted with a Gaussian function on the form:

$$f(x) = a \cdot \exp\left[-\frac{(x - b)^2}{c^2}\right], \tag{13}$$

where a, b, and c are fitting parameters. The FWHM was calculated from the parameter c, according to:

$$\text{FWHM} = 2\sqrt{\ln(2)} \cdot c \tag{14}$$

and is reported with a 95% confidence interval.

Flexible-meccano

We have used the program Flexible-meccano [19] with default settings to generate a pool of 10,000 possible polypeptide backbones by randomly selecting specific amino-acid conformations from a library of non-secondary structural elements of high-resolution X-ray crystallographic structures.

Experiments

Sample preparation

Statherin was purchased from Genemed Synthesis, Inc.. A 20 mM Tris [>99.9%, CAS (77-86-1); Saveen Werner AB] buffer with 150 mM NaCl [reagent grade, CAS (7647-14-5); Sharlau] was prepared with Milli-Q water, and the pH was set to 8.1 by dropwise addition of 1 M HCl, and thereafter, it was filtered through a hydrophilic polypropylene 0.2 μm membrane (Pall Corporation). The protein powder was dissolved in buffer by a small addition of NaOH to increase the pH, since the protein powder contained trifluoroacetate. A concentrating cell (Vivaspin 2, 2000 MWCO, Prod. No. VS02H92; Sartorius, Cambridge, United Kingdom) was used to remove low-molecular-weight impurities. The sample was rinsed with buffer corresponding to 30 times the sample volume, by centrifugation at 1600 rpm at 8°C. To ensure an exact background in the SAXS measurements, the sample was dialyzed (Slide-A-Lyzer Dialysis Cassette, 2000 MWCO, Prod. No. 66203; Thermo Scientific, Waltham, MA, USA) overnight at 6°C. Before the SAXS measurements, the sample was centrifuged at 14,000 rpm at 6°C for at least 2 h to remove aggregates. Thereafter, it was diluted to a concentration series, and the protein concentration was determined with a nanodrop spectrometer at the beamline using $\lambda = 280$ nm and $\varepsilon = 8740\ \mathrm{M}^{-1}\ \mathrm{cm}^{-1}$. The samples were centrifuged in small PCR tubes imminent to the SAXS measurements to remove any bubbles.

SAXS measurements

SAXS experiments were performed at BM29, ESRF-Grenoble, France. The incident beam wavelength was 0.99 Å, and the distance between sample and detector (PILATUS 1M) was set to 2867 mm, giving the scattering vector 0.0039–0.49 Å^{-1}. The scattering vector, q, is defined as $q = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident beam. Several successive frames of the scattering from the samples were recorded with a 0.5-s exposure time. The scattering from the pure solvent, which was measured before and after each sample for the same exposure times, was subtracted from the sample scattering. All measurements were performed at 20°C, and I_0 was converted to absolute scale by measuring the scattering of water. SAXS data were measured either after passing through a size exclusion chromatography (SEC) column or within a flowing capillary. For the inline SEC-SAXS, 5 mg/mL protein was injected through a 100-μL loop into a Superdex 75 10/300 GL column (GE Healthcare), equilibrated in 20 mM Tris, with 150 mM NaCl and a pH of 8.1. During SEC-SAXS, data were collected with a 1 s exposure time.

SAXS analysis

The SAXS and SEC-SAXS data were extracted and processed using PRIMUS [38] and ScÅtter (available at www.bioisis.net), respectively. Special attention was paid to radiation damage by comparing the

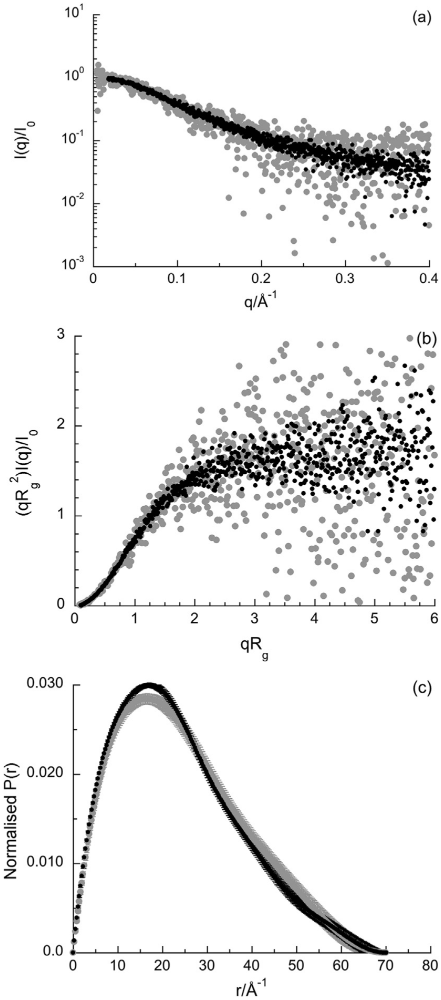


Fig. 14. SAXS data obtained for Statherin at 20 mM Tris and 150 mM NaCl (pH 8.1) at BM29, ESRF-Grenoble, France. Form factor (a), dimensionless Kratky plot (b), and pair distance distribution function, P(r) (c). The black circles correspond to data obtained from SEC in combination with SAXS, and the gray circles refer to continuous flow SAXS. If the precision is not visible, it is within the size of the data marker.

successive frames prior to background subtraction, and any affected data were rejected from further analysis. The form factor was obtained at the protein concentration 0.24 mg/mL, as shown in Fig. 14. From the pair distance distribution, P(r), the radius of gyration, R_g, was determined to be 19.8 ± 0.6 Å. The molecular weight was determined to be 5.29 kDa based on I_0 obtained from P(r). This is in good agreement with the theoretical molecular weight of 5.38 kDa, confirming that monomeric Statherin was obtained. The scattering curve from the peak in SEC-SAXS, also presented in Fig. 14, is in excellent agreement with the curve measured at 0.24 mg/mL, and R_g obtained from P(r) was determined to be 19.3 ± 0.2 Å. Hence, it is consistent with the measurement at 0.24 mg/mL. Since the protein concentration in the eluent from the SEC column was unknown, no molecular weight was obtained. However, due to the perfect agreement between the data obtained from SEC-SAXS and measured at 0.24 mg/mL, the less noisy SEC-SAXS data were used for comparison with simulations.

Acknowledgment

We are grateful to Dr. Mark Tully at the European Synchrotron Radiation Facility (ESRF), Grenoble, for providing assistance in using beamline BM29. This study was supported by the Science Faculty project grant program for research with neutrons and synchrotron light (Lund University Strategic funds for MAX-IV and European Spallation Source). The simulations were performed on resources provided by the Swedish National Infrastructure for Computing at the Center for scientific and technical computing at Lund University (LUNARC). We are also grateful to Boze *et al*. [15] for sharing their IB5 SAXS data.

Received 14 December 2017;
Received in revised form 3 March 2018;
Accepted 12 March 2018
Available online 21 March 2018

Keywords:
intrinsically disordered proteins;
coarse-grained modeling;
Monte Carlo simulations;
electrostatic interactions;
small-angle X-ray scattering

Abbreviations used:
IDPs, intrinsically disordered proteins; IDRs, Intrinsically disordered regions; SAXS, small-angle X-ray scattering; SARW, self-avoiding random walk; FCR, fraction of charged residues; FWHM, full width half maximum; NCPR, net charge per residue.

References

[1] A.K. Dunker, et al., Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.

[2] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (2002) 527–533.

[3] J.T. Liu, J.R. Faeder, C.J. Camacho, Toward a quantitative theory of intrinsically disordered proteins and their function, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19819–19823.

[4] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (2004) 635–645.

[5] M. Kjaergaard, A.-B. Norholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen, B.B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II? Protein Sci. 19 (2010) 1555–1564.

[6] R.B. Best, W. Zheng, J. Mittal, Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association, J. Chem. Theory Comput. 10 (2014) 5113–5124.

[7] J. Henriques, C. Cragnell, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment, J. Chem. Theory Comput. 11 (2015) 3420–3431.

[8] J. Henriques, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models, J. Chem. Theory Comput. 12 (2016) 3407.

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

[10] S. Rauscher, V. Gapsys, M.J. Gajda, M. Zweckstetter, B.L. de Groot, H. Grubmueller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment, J. Chem. Theory Comput. 11 (2015) 5513–5524.

[11] D.A. McQuarrie, Statistical Mechanics, 1st ed. University Science Books, Sausalito, Califonia, 2000.

[12] C. Cragnell, D. Durand, B. Cabane, M. Skepo, Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS, Proteins Struct. Funct. Bioinf. 84 (2016) 777–791.

[13] E.W. Martin, A.S. Holehouse, C.R. Grace, A. Hughes, R.V. Pappu, T. Mittag, Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation, J. Am. Chem. Soc. 138 (2016) 15323–15335.

[14] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A.K. Dunker, et al., pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins, Nucleic Acids Res. 42 (2014) D326–D335.

[15] H. Boze, T. Marlin, D. Durand, J. Perez, A. Vernhet, F. Canon, et al., Proline-rich salivary proteins have extended conformations, Biophys. J. 99 (2010) 656–665.

[16] S. Jephthah, J. Henriques, C. Cragnell, S. Puri, M. Edgerton, M. Skepo, Structural characterization of histatin 5–spermidine conjugates: a combined experimental and theoretical study, J. Chem. Inf. Model. 57 (2017) 1330–1341.

[17] H.A. Bruce, D. Du, D. Matak-Vinkovic, K.J. Bandyra, R.W. Broadhurst, E. Martin, et al., Analysis of the natively

unstructured RNA/protein-recognition core in the *Escherichia coli* RNA degradosome and its interactions with regulatory RNA/Hfq complexes, Nucleic Acids Res. 46 (2018) 387–402.

[18] D.P. O'Brien, B. Hernandez, D. Durand, V. Hourdel, A.-C. Sotomayor-Perez, P. Vachette, et al., Structural models of intrinsically disordered and calcium-bound folded states of a protein adapted for secretion, Sci. Rep. 5 (2015).

[19] V. Ozenne, F. Bauer, L. Salmon, J.-R. Huang, M.R. Jensen, S. Segard, et al., Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables, Bioinformatics 28 (2012) 1463–1470.

[20] R.K. Das, K.M. Ruff, R.V. Pappu, Relating sequence encoded information to form and function of intrinsically disordered proteins, Curr. Opin. Struct. Biol. 32 (2015) 102–112.

[21] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol 157 (1982) 105–132.

[22] H. Lee, A.H. de Vries, S.-J. Marrink, R.W. Pastor, A coarse-grained model for polyethylene oxide and polyethylene glycol: conformation and hydrodynamics, J. Phys. Chem. B 113 (2009) 13186–13194.

[23] Flory, Principles of Polymer Chemistry, Cornell Univ. Press, Ithaca, NY, 1953.

[24] P. Bernado, M. Blackledge, A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering, Biophys. J. 97 (2009) 2839–2845.

[25] A. Borgia, W. Zheng, K. Buholzer, M.B. Borgia, A. Schueler, H. Hofmann, et al., Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods, J. Am. Chem. Soc. 138 (2016) 11714–11726.

[26] G. Fuertes, N. Banterlea, K.M. Ruff, A. Chowdhury, D. Mercadante, C. Koehler, et al., Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E6342–E6351.

[27] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, B. Schuler, Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 16155–16160.

[28] J.E. Kohn, I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, et al., Random-coil behavior and the dimensions of chemically unfolded proteins, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 12491–12496.

[29] I.S. Millet, S. Doniach, K.W. Plaxco, Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins, Unfolded Proteins 62 (2002) 241–262.

[30] D.K. Wilkins, S.B. Grimshaw, V. Receveur, C.M. Dobson, J.A. Jones, L.J. Smith, Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques, Biochemistry 38 (1999) 16424–16431.

[31] J.A. Riback, M.A. Bowman, A.M. Zmyslowski, C.R. Knoverek, J.M. Jumper, J.R. Hinshaw, et al., Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water, Science 358 (2017) 238–241.

[32] D.G. Angelescu, P. Linse, Branched-linear polyion complexes investigated by Monte Carlo simulations, Soft Matter 10 (2014) 6047–6058.

[33] C.D. Andrew, J. Warwicker, G.R. Jones, A.J. Doig, Effect of phosphorylation on alpha-helix stability as a function of position, Biochemistry 41 (2002) 1897–1905.

[34] M. Zachariou, I. Traverso, L. Spiccia, M.T.W. Hearn, Potentiometric investigations into the acid-base and metal ion binding properties of immobilized metal ion affinity chromatographic (IMAC) adsorbents, J. Phys. Chem. 100 (1996) 12680–12690.

[35] B.D. Halligan, V. Ruotti, W. Jin, S. Laffoon, S.N. Twigger, E. A. Dratz, ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels, Nucleic Acids Res. 32 (2004) W638-W644.

[36] K. Binder, Monte Carlo and Molecular Dynamics Simulations in Polymer Science, Oxford University Press, New York, 1995.

[37] J. Rescic, P. Linse, MOLSIM: a modular molecular simulation software, J. Comput. Chem. 36 (2015) 1259–1274.

[38] P.V. Konarev, V.V. Volkov, A.V. Sokolova, M.H.J. Koch, D.I. Svergun, PRIMUS: a Windows PC-based system for small-angle scattering data analysis, J. Appl. Crystallogr. 36 (2003) 1277–1282.

[39] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, et al., Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase, Structure 18 (2010) 494–506.

Paper II

Assessing the Intricate Balance of Intermolecular Interactions upon Self-Association of Intrinsically Disordered Proteins

Ellen Rieloff[1], **Mark D. Tully**[2] **and Marie Skepö**[1]

*1 - **Theoretical Chemistry,** Lund University, POB 124, SE-221 00 Lund, Sweden*
*2 - **European Synchrotron Radiation Facility (ESRF),** Grenoble, France*

Correspondence to Ellen Rieloff and Marie Skepö: *ellen.rieloff@teokem.lu.se, marie.skepo@teokem.lu.se*

https://doi.org/10.1016/j.jmb.2018.11.027
Edited by Monika Fuxreiter

Abstract

Attractive interactions between intrinsically disordered proteins can be crucial for the functionality or, on the contrary, lead to the formation of harmful aggregates. For obtaining a molecular understanding of intrinsically disordered proteins and their interactions, computer simulations have proven to be a valuable complement to experiments. In this study, we present a coarse-grained model and its applications to a system dominated by attractive interactions, namely, the self-association of the saliva protein Statherin. SAXS experiments show that Statherin self-associates with increased protein concentration, and that both an increased temperature and a lower ionic strength decrease the size of the formed complexes. The model captures the observed trends and provides insight into the size distribution. Hydrophobic interaction is considered to be the major driving force of the self-association, while electrostatic repulsion represses the growth. In addition, the model suggests that the decrease of association number with increased temperature is of entropic origin.

Introduction

Intrinsically disordered proteins (IDPs) are characterized by a lack of stable tertiary structure under physiological conditions *in vitro* [1,2] and hence are best described by conformational ensembles [3,4]. Bioinformatic studies have led to the conclusion that 10%–20% of the eukaryotic proteins are intrinsically disordered, and even more proteins contain intrinsically disordered regions (IDRs) [5–8]. It has also been established that IDPs and IDRs are involved in many biological processes and diseases, and that the lack of folded structure is related to their functions [7,9].

Attractive interactions between IDPs can lead to the formation of aggregates, which in the case of diseases such as Parkinson's disease and Alzheimer's disease is harmful [10]. IDP attractions can also be fundamental for a desired outcome, such as in the formation of proteinaceous membrane-less organelles [11–14], which are condensed liquid droplets often enriched in IDPs and IDRs and commonly found in the cell cytoplasm and nucleus [15]. Various pieces of evidence suggest that liquid–liquid phase separation is a driving force for the formation of some proteinaceous membrane-less organelles [11–14], and that the phase separation itself is driven by weak multivalent interactions between disordered proteins [16,17].

For understanding IDPs and their interactions, computer simulations are a useful complement to experiments [18,19]. There have been considerable advances regarding atomistic simulations of IDPs, where development and justification of force fields and water models have been validated against experimental results [20–24]. The full-atom approach and explicit water treatment in atomistic simulations are great advantages for gaining a molecular understanding, however, atomistic simulations are computationally demanding, both regarding execution time and data storage. Hence, this poses limitations on the accessible timescale and system size, and therefore, a coarse-grained approach is a more viable option for studying complex systems, such as the examples above. Recently, a coarse-grained model based on the primitive model, in combination with Monte Carlo simulations, has proven capable of capturing bulk properties at dilute conditions for a range of IDPs [25]. We aim to develop this model to also account for more complex systems, and first is the investigation of a model system dominated by intermolecular attractions, namely, the self-association of the saliva protein Statherin. Statherin has a distinct amphiphilic character in its primary sequence, shown in Fig. 1. Almost all charges are located in the N-terminal part, starting with a block of negative charges, followed by a block of positive charges. From the hydropathy values in the Kyte–Doolittle scale [26], it is shown that overall the hydropathy is rather low, which is typical for IDPs. However, residues 15–43 contain seven tyrosines, whose aromatic side chains have been established to be of importance for liquid–liquid phase separation [27,28]. Statherin also consists of 16% proline residues, which are denoted as "disorder-promoting" [29].

In this work, Statherin is characterized experimentally at monomeric conditions through the use of small-angle X-ray scattering (SAXS) and circular dichroism (CD), and at self-associating conditions through SAXS experiments and simulations. The simulation model is validated against the experiments and is demonstrated to be useful for describing polydispersity and the interplay between electrostatics, hydrophobic interactions, and entropy in the self-association process.

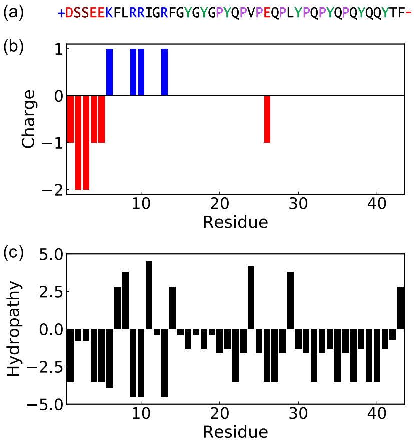


Fig. 1. (a) Amino acid sequence of Statherin with the charge distribution at pH 8 and certain amino acids highlighted. Positive residues are marked in blue, negative in red, phosphorylated serines with the charge −2*e* in dark red, and prolines in lilac and tyrosines in green. (b) Charge distribution and (c) hydropathy values using the Kyte–Doolittle scale, where −4.5 is the most hydrophilic and +4.5 is the most hydrophobic [26].

Results and Discussion

The experimental results for Statherin at monomeric conditions are presented first, followed by the self-association studied both experimentally and by Monte Carlo simulations.

Monomeric behavior

In Fig. 2a–c, data for monomeric Statherin obtained by SAXS coupled with size-exclusion chromatography (SEC' taken from Ref. [25]) is presented. From regular SAXS measurements at low protein concentration (0.24 mg/mL), the molecular weight was determined to be 5.29 kDa, based on the forward scattering, I_0, obtained from the pair distance distribution function, $P(r)$ [25]. This is in good agreement with the theoretical molecular weight of 5.38 kDa, confirming monomeric conditions. As seen in Fig. 2a, Statherin shows the typical featureless scattering profile of an IDP. The IDP character is also verified by the dimensionless Kratky plot in Fig. 2b, where the profile has an uprise slope and reaches a plateau at higher q values, typical for flexible chains. In addition, the CD data presented in Fig. 2d confirm a random coil behavior with some presence of secondary structure. The global minimum is located at 205 nm, which is slightly higher than the usual 198 nm for random coils; however, it is typical for poly-proline II (PPII) structure. The shallow minimum close to 222 nm might suggest a small presence of α-helix. Several studies of Statherin with CD or NMR have suggested that the charged N-terminal has a propensity for forming α-helix and that a part of the middle adopt PPII structure. Nevertheless, the overall structure is still disordered in aqueous solution [30–34]. Fig. 2d also shows that there are no large differences in structure due to salt concentration.

The radius of gyration for monomeric Statherin in 150 mM NaCl has been reported as 19.3 ± 0.2 Å, based on the $P(r)$ presented in Fig. 2c [25]. With urea, the radius of gyration is increased to 22.1 ± 0.2 Å for 4 M urea and to 23.7 ± 0.3 Å for 8 M urea. The dimensionless Kratky plot, shown in Fig. 3a, also indicates an increase in stiffness when urea is added. From CD measurements it is seen that the mean residue ellipticity ($[\theta]_{\mathrm{MRW}}$) at 228 nm, presented in Fig. 3b and c, increases linearly with increased urea concentration and also becomes positive at high urea concentrations. This corresponds to an increase of PPII content, in agreement with the study by Whittington *et al.* [35], reporting that urea promotes PPII formation. PPII conformation is more extended than both random coil and α-helix; hence, this explains the changes observed in the SAXS measurements.

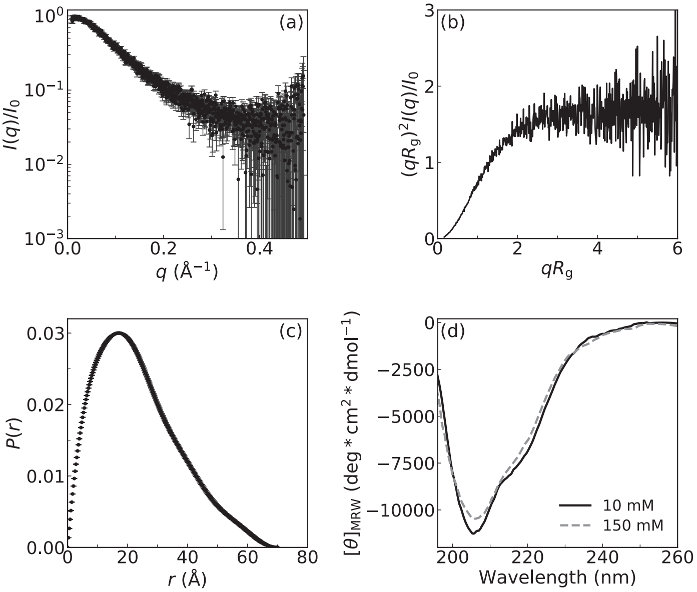


Fig. 2. SAXS data for Statherin obtained by SEC-SAXS, at 150 mM NaCl and 20 mM Tris buffer with pH 8, from Ref. [25]. (a) Form factor, (b) dimensionless Kratky plot, and (c) pair distance distribution function. (d) CD spectra for Statherin in 10 and 150 mM NaF and 20 mM phosphate buffer (pH 8) with a protein concentration of 0.11 and 0.13 mg/mL, respectively, measured at 20 °C.

Temperature also induces changes in secondary structure. With increased temperature, the $[\theta]_{MRW}$ increases at 205 nm and decreases at 228 nm, as shown in Fig. 4, suggesting a loss of PPII as described by Kjaergaard *et al*. [36] for other IDPs. The loss of PPII appears rather proportional to temperature.

Self-association

Experimental results

With increased protein concentration, Statherin self-associates into complexes, which is evident from an increase in forward scattering. The average number of proteins per complex was determined from the forward scattering and is presented against the protein concentration in Fig. 5a for the reference system with 150 mM NaCl. Panels b and d in the same figure present corresponding data from simulations and will be discussed in the next section. The growth is linear with respect to concentration up to 10 mg/mL, and afterward, the slope decreases, which might suggest a maximum size of the Statherin complex. Likewise, the radius of gyration follows the same trend, although a plateau is reached earlier. However, a depression of the forward scattering at higher concentrations due to a structure factor cannot be ruled out, and therefore, the high concentration data should be interpreted with care. Especially since, at 24 mg/mL and higher concentrations, inter-particle interference is visible in the $P(r)$ as a decrease below zero at long distances. The scattering curves, Guinier plots, and I_0 and radius of gyration determined by both Guinier and $P(r)$ are provided in Supplemental information.

The Kratky plot in Fig. 5c shows a transition from flexible chain behavior to more globular when the complexes are formed. The complexes are also more spherical in shape than the free proteins, which is evident from the pair distance distribution function presented in Fig. 6, plotted to enhance the differences compared to a sphere.

Since urea weakens hydrophobic interactions [37], the effect of urea on the Statherin complexes was studied. With 8 M urea, no increase in forward scattering was observed even when reaching 32 mg/mL in protein concentration. The only effect observed was a lowering of the forward scattering due to a structure factor emerging. This indeed suggests hydrophobic interactions as a driving force for the self-association in Statherin. With 4 M urea, it was a downshift at intermediate q when going from 2 to 4 mg/mL and that continued for even higher protein concentrations (data not shown). This in

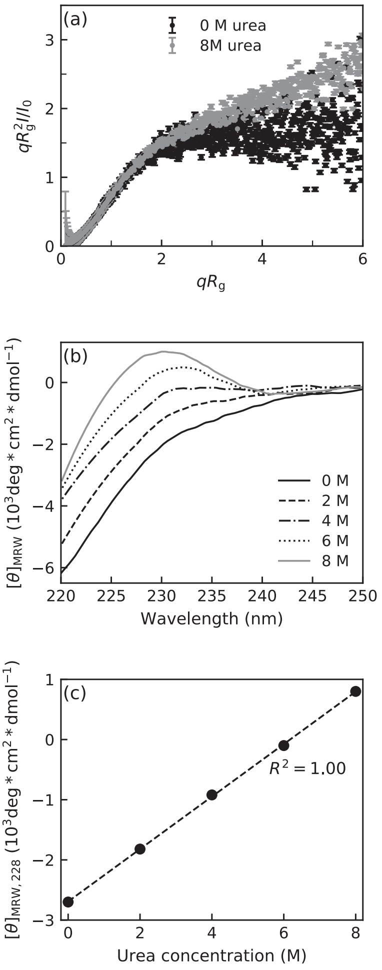


Fig. 3. Effect of urea. (a) Dimensionless Kratky plot for Statherin at 150 mM NaCl measured by SEC-SAXS and with 8 M urea measured by SAXS at a protein concentration of 4 mg/mL, (b) CD spectra and (c) mean residue ellipticity at 228 nm for Statherin (0.12–0.14 mg/mL) *versus* urea concentration, obtained from CD measurements at 20 °C and pH 8.

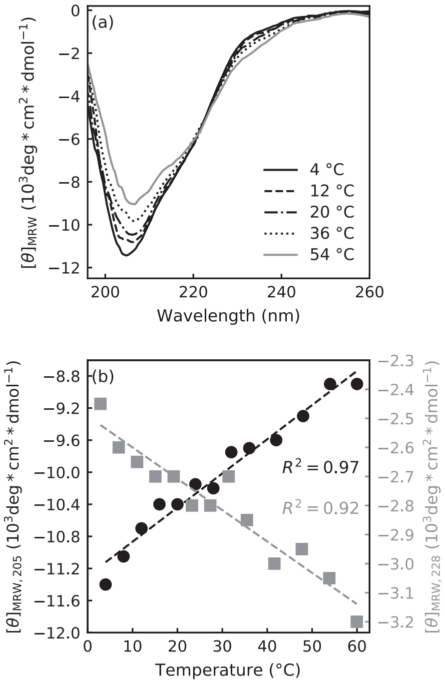


Fig. 4. Temperature dependence of monomeric Statherin (0.13 mg/mL) with 150 mM NaF in 20 mM phosphate buffer at pH 8. (a) CD spectra and (b) mean residue ellipticity at 205 nm (black circles) and 228 nm (gray squares).

combination with a decrease in slope in the Kratky plot with increasing concentration suggests that there are still complexes forming in 4 M urea. For surfactants, both the critical micelle concentration and the micelle size have been reported to change with the concentration of urea [38–40].

Self-association has been observed no matter the salt concentration, which supports hydrophobic interactions being the major driving force. However, the average association number appears to increase with increased ionic strength, as presented in Fig. 7a. Due to the possibility of structure factor influence on the scattering data at lower ionic strength, the effect of electrostatic interactions is further discussed within the framework of the simulations (data presented in Fig. 7b).

Changing the temperature also affects the self-association, as shown by a decrease in association number with increased temperature in Fig. 8. The average radius of gyration follows the same trend (data not shown). The decrease of the association number with temperature has also been observed for surfactants with ionic or zwitterionic headgroups [41], while non-ionic surfactants have shown the opposite temperature dependence [41,42]. For the intrinsically disordered milk-protein β-casein, the association number increases with increased temperature at neutral pH [43], as for non-ionic surfactants. Although β-casein and Statherin have similar block structures, the overall hydrophobicity is higher in β-casein. Hence, it is not unreasonable that the temperature dependence is different.

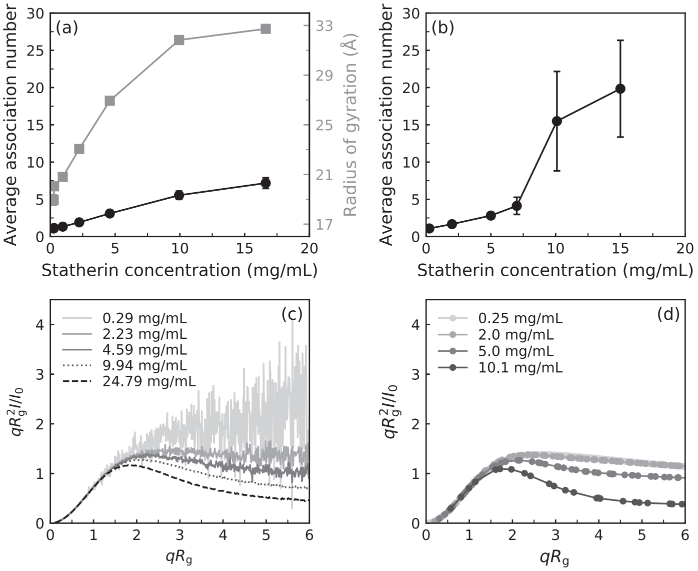


Fig. 5. (a) Average number of proteins per complex (black circles) and radius of gyration (gray squares) *versus* protein concentration determined from SAXS. (b) Average number of proteins per complex *versus* protein concentration from simulations. (c) Dimensionless Kratky plot from experiments. (d) Dimensionless Kratky plot from simulations. The data is reported for the reference system (experimental conditions: 20 mM Tris, 150 mM NaCl, pH 8, 20 °C; simulation conditions: 150 mM implicit salt, 20 °C). In panel a, the error bars on the association number represent a 10% uncertainty.

Simulation results

We have simulated the Statherin system using a modified version of the coarse-grained model presented in Ref. [25]. Therein it was shown that the coarse-grained model works well for Statherin at monomeric conditions. However, to capture the self-association, an additional attractive interaction is needed. We have implemented a short-ranged potential corresponding to 1.32 *kT* at closest contact between neutral amino acids, mimicking a smeared hydrophobic interaction, which causes the proteins to associate upon increased concentration. For the reference system, 150 mM salt, the simulation data follow the linear trend described in experimental data up to approximately 7 mg/mL, according to Fig. 5b. Then it deviates, by forming large complexes, which shall be interpreted as that the model is reliable only at lower protein concentrations. The model is able to capture the experimentally established transition to a more globular state with increased protein concentration in the Kratky plot, c.f. Fig. 5d and c, although the single chain is too compact due to the extra attraction. To capture the behavior at both monomeric conditions and higher protein concentrations, an angular potential can be included as well. However, since the goal with this model is to capture general trends, an exact matching with the experimental Statherin data is not important, and hence, the results of the model without further modifications are presented.

The simulations show that the complexes are polydisperse; see the complex size probability distribution in Fig. 9a. At 7 mg/mL and lower concentrations,

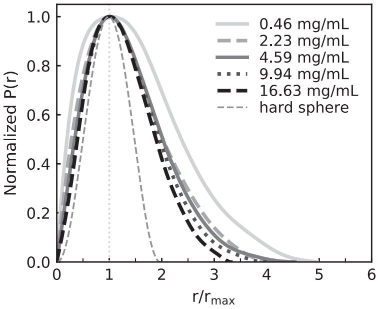
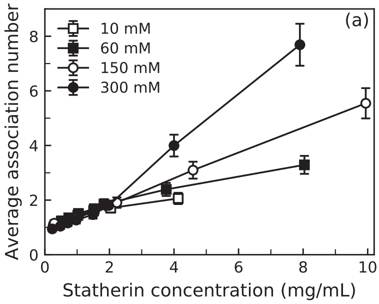


Fig. 6. Pair distance distribution function normalized to enhance deviations in shape from a homogeneous hard sphere, where r_{max} corresponds to the value of r where $P(r)$ has its maximum, for the reference system (20 mM Tris, 150 mM NaCl, pH 8, 20 °C).



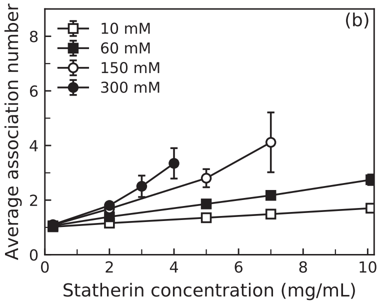


Fig. 7. Average association number determined (a) by SAXS and (b) from simulations, as a function of Statherin concentration for different concentrations of NaCl, at 20 °C. The error bars in panel a represent a 10% uncertainty.

the monomer is the dominating specie and the amount of the different species decreases with increasing size. The polydispersity and monomeric dominance is also evident from the snapshot in Fig. 9b, which furthermore suggests that it is the middle and C-terminal part that forms the core of the complex and that the charged N-terminal part is located on the surface of the complex.

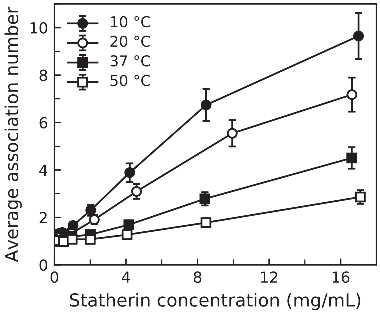


Fig. 8. Average number of proteins per complex determined by SAXS *versus* protein concentration at 150 mM NaCl for 10–50 °C. The error bars represent a 10% uncertainty. The data at 20 °C correspond to the data at 150 mM NaCl in Fig. 7a.

The contact probability between residues of different chains is presented in Fig. 9c and confirms indeed that it is the neutral amino acids that are mostly in contact with other chains. In Fig. 9d, the radial number density distribution from the complex center of mass is presented. It again confirms that the core consists of neutral residues. The negatively charged residue 26 is also part of the core of the complex. The other charged residues are located closer to the surface of the complex.

The experimental $P(r)$ in Fig. 5d shows that the complexes are more spherical than the monomers, due to the change with increasing concentration. However, the experiments only provide the average over all different complex sizes. In the simulations, we have calculated the principal moments of the gyration tensor and from that the asphericity for the complexes of different sizes. It indeed confirms that the monomers are not spherical, having an asphericity value of 0.41. The asphericity decreases with increasing association number until six, where it stabilizes around 0.13 also for larger complexes. If the asphericity is less than 0.1, the object is normally considered spherical [44]. The decrease in asphericity agrees with the experimental results and furthermore shows that the complexes are close to the spherical limit. However, for complexes consisting of seven protein chains, $\langle R_1^2 \rangle$, $\langle R_2^2 \rangle$ and $\langle R_3^2 \rangle$ were 323.5 ± 7.1 Å^2, 158.2 ± 1.2 Å^2, and 91.1 ± 0.5 Å^2, respectively, showing that the instantaneous shapes of the complexes are still not spherical.

The increase of size of the complexes with increased ionic strength observed in SAXS experiments is also captured by the simulations, as seen in Fig. 7b, even if the effect is slightly overestimated compared to experiments (Fig. 7a). This confirms that although the hydrophobic interaction is the major driving force for self-association, electrostatic repulsion stabilizes the system and depresses the growth. To further investigate the electrostatic effect, we performed simulations without phosphorylated serines, which increases the net charge from −4 to 0. This shifts the complex size probability distribution toward larger sizes, depicted in Fig. 10. The overall contact probability also increases from 0.36 ± 0.03 with phosphorylated serines to 0.41 ± 0.01 without phosphorylations at a protein concentration of 2 mg/mL, while the contact profile remains similar in shape. This demonstrates that phosphorylations indeed affect the electrostatic interactions and that it is of importance for the self-association.

Another mutation that illustrates the importance of electrostatics is the point mutation of residue 26, glutamic acid, changing the negatively charged residue located in the middle of the neutral block to a neutral residue. Already in a simulation at 2 mg/mL, the majority of the chains join in one large complex, while for comparison, the reference system rarely exhibits complexes larger than tetramers at the same

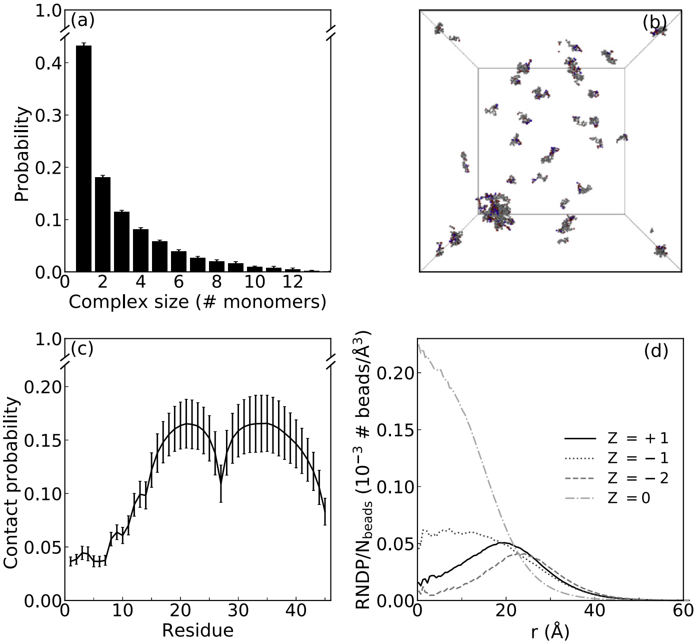


Fig. 9. Simulation data at 5 mg/mL with 150 mM implicit salt. (a) Complex size probability distribution. (b) Snapshot with excluded counterions, where gray beads represent neutral residues, red beads represent negatively charged residues, and blue beads represent positively charged residues. (c) Chain contact probability profile. (d) Radial number density for different bead types, normalized by the number of beads of each type in the protein, as a function of distance from the core center of mass, for complexes consisting of seven proteins. Z represents the charge of each bead type.

concentration. This shows that specific residues can make a great difference for the self-association (results not shown).

With increased temperature, the average association number, displayed in Fig. 11, decreases, again in accordance with experimental results. Since Statherin has a net charge of −4*e*, the overall electrostatic interaction is repulsive. Increased temperature enhances electrostatic interactions, and hence, it would counteract self-association by enhancing the net electrostatic repulsion between Statherin monomers. In addition, the effect of entropy, also opposing self-association, increases with temperature as well. Note that the hydrophobic interaction is regarded temperature-independent in this model. Simulations of the Statherin system without charges at a concentration of 4 mg/mL show a decrease in average association number between 20 and 50 °C, from 3.06 ± 0.63 to 1.39 ± 0.01, compared to 2.24 ± 0.15 to 1.40 ± 0.01 for the same system with charges. This suggests entropy as the main contribution to the temperature effect.

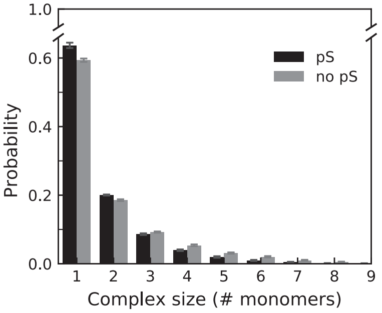


Fig. 10. Complex size probability distribution for 2 mg/mL Statherin with and without phosphorylated serines at 150 mM ionic strength.

Temperature also affects the structure of the complexes. Overall, the asphericity increases as a function of temperature for complexes of the same size, as seen in Fig. 11b. In addition, the radius of gyration also shows the same trend, for example, for complexes of seven proteins, the R_g goes from 22.8 ± 0.1 to 29.8 ± 0.2 Å when temperature changes from 15 to 50 °C. These changes reflect an

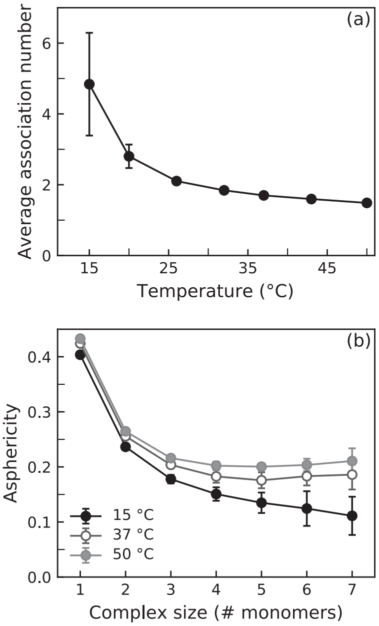


Fig. 11. (a) Average association number as a function of temperature at 5 mg/mL. (b) Asphericity *versus* association number at 15, 37 and 50 °C.

increased flexibility in the complexes, which is expected due to the entropy increase. Although it was shown in the monomeric section that the structure of the individual protein chain changes upon temperature increase, it is expected to be of minor importance for the self-association process, due to the model capturing the trends without including such detail.

Model limitations and improvements

From the simulations, it is apparent that the model breaks down at higher concentrations. The exact concentration depends on the conditions, especially temperature and ionic strength. At the lower-salt concentrations (10 and 60 mM), no breakdown is observed even at 20 mg/mL. The breakdown can be connected with the implicit treatment of salt, since simulations with 150 mM explicit salt and 20 mg/mL protein or more still give an average size less than 10 chains/complex. Hence, an explicit treatment of electrostatics is suggested to provide better results, although at a high computational cost. In the model, the hydrophobic interaction, mimicking the effect of both the enthalpic contribution and the entropic effect on the water molecules, is regarded temperature independent. Including temperature dependence would change the exact values to a certain extent, although the trend would remain. Hence, it would not affect the conclusion that entropy in the system is the largest contributor to the temperature effect for this protein.

Conclusions

A modified version of the coarse-grained model in Ref. [25] have been shown capable to describe the Statherin complexes at lower concentration and provide extra insight regarding the structure of the complexes, as well as aiding in explaining the effect of external conditions on the self-association, in terms of a balance between different interactions and entropy. The findings are summarized in Fig. 12. Hydrophobic interaction is shown to be the major driving force for the self-association, due to urea inhibiting complex formation. The size decrease as a result of increased temperature is regarded as an entropic effect, while electrostatic interactions were

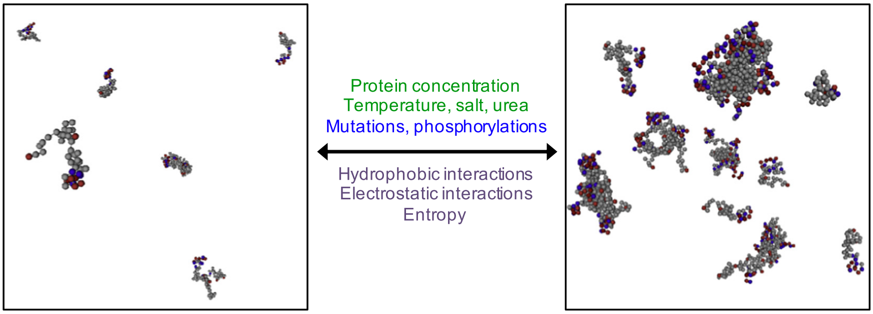


Fig. 12. Summary of what was shown to affect the Statherin association state. External factors are printed in green, chain characteristics in blue, and energetic and entropic factors in purple. In the snapshots, gray beads represent neutral residues; blue, positively charged residues; and red, negatively charged residues. The phosphorylated serines are marked in dark red. Counterions are omitted for clarity.

still shown to be of importance by balancing the hydrophobic attraction. In addition, it was demonstrated that mutations affecting the charge distribution can have a major effect on the self-association.

The self-association of Statherin is only one example of an IDP system dominated by intermolecular attractions; however, the similarities to micelle formation suggest that the established interactions are common for many systems, although with varying balance. It is therefore of interest to apply this model to other interacting IDPs in the future, as well as to continue the development for studies of systems with a higher complexity. Computational studies of IDP systems are advantageous in that it allows for separation of different contributions and a faster screening of mutations. In combination with experiments, it opens up for a deeper understanding of the function and behavior of IDPs.

Methods and Model

SAXS

Sample preparation

The buffers, all containing 20 mM Tris [>99.9%, CAS (77-86-1); Saveen Werner AB], and varying concentrations of NaCl [reagent grade, CAS (7647-14-5); Sharlau] and urea [ReagentPlus ≥99.5%, CAS (57-13-6); Sigma-Aldrich] were prepared with Milli-Q water, and by dropwise addition of 1 M HCl, the pH was set at room temperature to correspond to 8.1 at the measuring temperature. Thereafter, the buffers were filtered through a hydrophilic polypropylene 0.2 μm membrane (Pall Corporation).The Statherin powder (purchased from Genemed Synthesis, Inc.) was dissolved in buffer with a small addition of NaOH to increase the pH, since the protein powder contained trifluoroacetate. Concentrating cells (Vivaspin 2, 2000 MWCO, Prod. No. VS02H92; Sartorius, Cambridge, United Kingdom) were used to remove low-molecular-weight impurities. The samples were rinsed with buffer corresponding to 30 times the sample volume, by centrifugation at 358*g* at 8 °C. To ensure an exact background in the SAXS measurements, the samples were dialyzed (Slide-A-Lyzer Dialysis Cassette, 2000 MWCO, Prod. No. 66203 or Slide-A-Lyzer MINI Dialysis Unit, 2000 MWCO, Prod. No. 69580; Thermo Scientific, USA) overnight at 6 °C. Before the SAXS measurements, the samples were centrifuged at 18,400*g* at 6 °C for at least 2 h to remove impurities. Thereafter, they were diluted to a concentration series, and the protein concentration was determined with a nanodrop spectrometer using $\lambda = 280$ nm and $\varepsilon = 8740\ \mathrm{M}^{-1}\ \mathrm{cm}^{-1}$. The samples were centrifuged in small PCR tubes imminent to the SAXS measurements to remove any bubbles.

Measurements and analysis

SAXS experiments were performed at BM29, ESRF-Grenoble, France. The incident beam wavelength was 0.99 Å, and the distance between sample and detector (PILATUS 1M) was set to 2867 mm, giving the scattering vector 0.0039 – 0.49 Å^{-1}. The scattering vector, q, is defined as $q = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident beam. Several successive frames of the scattering from the samples were recorded with an exposure time of 0.5 or 1 s, depending on concentration and system. The scattering from the pure solvent, which was measured before and after each sample for the same exposure times, was subtracted from the sample scattering. Measurements were performed at 10, 20, 37 and 50 °C at 150 mM NaCl, and the forward scattering, I_0, was converted to absolute scale by water calibration. At 20 °C measurements were also performed for 10, 60 and 300 mM NaCl and 4 and 8 M urea. The data were processed and analyzed using the ATSAS package [45]. Special attention was paid to radiation damage by comparing the successive frames prior to background subtraction, and any affected data were rejected from further analysis. Both I_0 and R_g were determined from $P(r)$, although the Guinier approach was also used for comparison. The molecular weight used for calculating the association number was determined from I_0 (see Supplemental information). Considering standard uncertainties of the used values, the uncertainty of the association number can be estimated as approximately 10% [43,46].

For a description of the SEC inline with SAXS, used for obtaining the form factor of monomeric Statherin, we refer to Ref. [25].

CD

Protein was dissolved in and purified with 20 mM phosphate buffer (sodium phosphate dibasic dihydrate [Reag. Ph. Eur., CAS (10028-24-7); Sigma-Aldrich] and sodium phosphate monobasic monohydrate [ACS reagent, CAS (10049-21-5); Sigma-Aldrich]) at pH 8, using a concentrating cell, as described for the SAXS samples. The protein was diluted to approximately 0.13 mg/mL using 20 mM phosphate buffer with 10 or 150 mM NaF [≥99%, CAS (7681-49-4); Sigma-Aldrich] and for the 150 mM NaF with 0–8 M urea [ReagentPlus ≥99.5%, CAS (57–13-6); Sigma-Aldrich]. The samples were filtered using a 0.22-μm Millex–GV filter (Merk Millipore Ltd). CD spectra between 190 and 260 nm at temperatures 4 – 60 °C were recorded on a JASCO J-715 instrument with a PTC-348WI Peltier type cell holder for temperature control, averaging over three spectra for each sample, using a quartz cuvette with a 1-mm path length (HellmaAnalytics) and 20-nm/min scanning speed, 2-s response time, 1-nm band width, and 100-mdeg sensitivity. At 20 °C, further measurements were performed for samples with 150 mM NaF and 2–8 M urea. The ellipticity reported is the mean residue ellipticity, defined as

$$[\theta]_{\mathrm{MRW}} = \theta \cdot \mathrm{MRW}/(10 \cdot d \cdot c), \tag{1}$$

where θ is the observed ellipticity (mdeg), d the path length of the cell (cm), and c the protein concentration (mg/mL). The mean residue weight, MRW, is the molecular weight (Da) divided by the number of peptide bonds. The spectra were smoothed using a Savitzky–Golay filter. The effect of the Savitzky–Golay filter is presented in Fig. S4 in Supplemental information.

Coarse-grained model

We have employed a coarse-grained model in which each amino acid is modeled as a hard sphere, further described in Ref. [25]. For the inclusion of hydrophobic interaction, a short-ranged potential is added to the model:

$$U_{\mathrm{hphob}} = -\sum_{\mathrm{neutral}} \frac{\varepsilon_{\mathrm{hphob}}}{r_{ij}^{6}} \tag{2}$$

where the summation extends over all neutral amino acids, $r_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ is the center-to-center distance between two beads and **R** refers to the coordinate vector. $\varepsilon_{\mathrm{hphob}}$ is $1.32 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of 1.32 kT at closest contact, determined by comparing the average complex size with experimental results on the reference system.

Simulation aspects

The equilibrium properties of the model systems were obtained by Metropolis Monte Carlo simulations in the canonical (NVT) ensemble, utilizing the simulation package Molsim [47], version 4.8.8. Forty-five protein chains were enclosed in a cubic box of varying volume, dependent on the protein concentration. Periodic boundary conditions were applied in all directions. The long-ranged Coulomb interactions were truncated using the minimum image convention.

To accelerate the examination of the configurational space, five different types of displacements were allowed: (i) translational displacement of a single bead, (ii) pivot rotation [48,49], (iii) translation of the entire chain, (iv) slithering move [50], and (v) cluster displacements. Counterions were only moved individually by translation. The cluster displacement was performed as a translational displacement of the chain of a selected particle as well as all chains whose center of mass were less than 40 Å away from the selected particle. The cluster displacement was automatically rejected if the number of particles within the cluster changed, that is, if the displacement caused two clusters to merge. The probability of the different trial moves was weighted so that 80% of the trial moves were single bead displacements, 5% were pivot rotations, 5% were chain displacements, 3% were slithering moves, and 7% were cluster moves. Initially, the proteins were randomly placed in the box and an equilibrium simulation of typically $3 \cdot 10^5$ trial moves/bead was performed. The proceeding production run comprised at least 10^6 passes divided into subdivisions of 10^5 passes. To ensure accurately sampled simulations, the contact probability of each chain individually and the variations of contact number along the propagation of the simulation were analyzed (data not shown).

For all simulated quantities except the average association number, the reported uncertainty is one standard deviation of the mean. It is estimated from the deviation among the means of the subdivisions of the total number of MC passes, according to

$$\sigma^2(\langle x\rangle) = \frac{1}{n_s(n_s-1)} \sum_{s=1}^{n_s} \left(\langle x\rangle_s - \langle x\rangle\right)^2, \tag{3}$$

where $\langle x\rangle_s$ is the average of quantity x from one subdivision, $\langle x\rangle$ the average of x from the total simulation, and n_s the number of subdivisions. For the average association number, the reported uncertainty is the standard deviation of the means of all subdivisions.

Analyses

The calculation of the scattering profile from simulation is described in Ref. [25]. In the analyses of complexes, two chains were assigned to the same complex if the center-to-center distance between two beads in the two different chains was less than 5 Å. The same geometric condition was used for defining if a bead was in contact with another chain, which was the basis for monitoring the variations of contact number along the propagation, and calculating the contact probability for beads along the chain. Contact probability for the beads is defined as the number of passes in which the bead is in contact with at least one bead from another chain, divided by the total number of passes in the simulation. Similarly, contact probability for a chain is calculated as the number of passes in which the chain is in a complex divided by the total number of passes in the simulation and the overall contact probability is the average over all chains. The complex size probability distribution was calculated according to

$$P_n = \frac{n\langle N_n^{\mathrm{complex}}\rangle}{\sum_n n\langle N_n^{\mathrm{complex}}\rangle}, \tag{4}$$

where $\langle N_n^{\text{complex}} \rangle$ is the average number of complexes consisting of n chains, and $\sum_n n\langle N_n^{\text{complex}} \rangle$ is equal to the number of chains in the system, due to chain conservation. Note that P_n is weighted by the number of chains in a complex. The average association number was calculated from the complex size probability distribution, as

$$N_{\text{assoc}} = \sum_n nP_n. \tag{5}$$

The radial number density profile was calculated for each complex size and bead type individually. The radial number density at each distance is defined as the number of beads within a shell at that distance from the center-of-mass of the complex core, divided by the shell volume. The complex core was defined to consist of the beads 15–44 in each chain.

The shape of the complexes was quantified by the principal moments of the gyration tensor and the asphericity. The gyration tensor was defined as

$$S = \frac{1}{N}\begin{pmatrix} \sum_i^N X_i^2 & \sum_i^N X_iY_i & \sum_i^N X_iZ_i \\ \sum_i^N X_iY_i & \sum_i^N Y_i^2 & \sum_i^N Y_iZ_i \\ \sum_i^N X_iZ_i & \sum_i^N Y_iZ_i & \sum_i^N Z_i^2 \end{pmatrix}, \tag{6}$$

where $A_i = (a_i - a_{\text{com}})$ for $a = x$, y, z, and N is the number of beads in the complex. Transformation to a principal axis system such that

$$S = \operatorname{diag}\left(R_1^2, R_2^2, R_3^2\right) \tag{7}$$

diagonalizes S and $R_1^2 \geq R_2^2 \geq R_3^2$ are the eigenvalues of S, also called the principal moments of the gyration tensor. In the simulations, the ensemble averages of the eigenvalues were calculated for each complex size separately. The asphericity, defined as

$$a_s = \frac{\left(\langle R_1^2\rangle - \langle R_2^2\rangle\right)\left(\langle R_2^2\rangle - \langle R_3^2\rangle\right)\left(\langle R_3^2\rangle - \langle R_1^2\rangle\right)}{2\left(\langle R_1^2\rangle + \langle R_2^2\rangle + \langle R_3^2\rangle\right)^2}, \tag{8}$$

ranges between 0 for a perfect sphere and 1 for a rod.

Acknowledgments

The authors thank the European Synchrotron Radiation Facility (ESRF) for providing beamtime, Dr. Bart Van Laer at ESRF for providing assistance in using beamline BM29, and Magnus Bergvall's foundation for financial support. The study was supported by the Science Faculty project grant program for research with neutrons and synchrotron light (Lund University Strategic funds for MAX-IV and European Spallation Source). The simulations were performed on resources provided by the Swedish National Infrastructure for Computing at the Center for scientific and technical computing at Lund University (Lunarc).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.11.027.

Received 10 September 2018;
Received in revised form 12 November 2018;
Accepted 28 November 2018
Available online 7 December 2018

Keywords:
intrinsically disordered proteins;
SAXS;
self-association;
Monte Carlo simulations;
coarse-graining

Abbreviations used:
IDPs, intrinsically disordered proteins; IDRs, intrinsically disordered regions; SAXS, small-angle X-ray scattering; SEC, size-exclusion chromatography; PPII, poly-proline II.

References

[1] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (10) (2002) 527–533, https://doi.org/10.1016/S0968-0004(02)02169-2.

[2] A. Dunker, J. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (1) (2001) 26–59, https://doi.org/10.1016/S1093-3263(00)00138-8.

[3] A.K. Dunker, J. Gough, Sequences and topology: intrinsic disorder in the evolving universe of protein structure, Curr. Opin. Struct. Biol. 21 (3) (2011) 379–381, https://doi.org/10.1016/j.sbi.2011.04.002.

[4] J. Habchi, P. Tompa, S. Longhi, V.N. Uversky, Introducing protein intrinsic disorder, Chem. Rev. 114 (13) (2014) 6561–6588, https://doi.org/10.1021/cr400514h.

[5] A.K. Dunker, P. Romero, Z. Obradovic, E.C. Garner, C.J. Brown, Intrinsic protein disorder in complete genomes, Genome Inform. 11 (2000) 161–171, https://doi.org/10.11234/gi1990.11.161.

[6] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca, E. Garner, S. Guilliot, A. Dunker, Thousands of proteins likely

to have long disordered regions, Pac. Symp. Biocomput. 3 (1998) 437–448.

[7] J. Ward, J. Sodhi, L. McGuffin, B. Buxton, D. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (3) (2004) 635–645, https://doi.org/10.1016/j.jmb.2004.02.002.

[8] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2) (2012) 137–149, https://doi.org/10.1080/07391102.2012.675145.

[9] J. Liu, J.R. Faeder, C.J. Camacho, Toward a quantitative theory of intrinsically disordered proteins and their function, Proc. Natl. Acad. Sci. U. S. A. 106 (47) (2009) 19819–19823, https://doi.org/10.1073/pnas.0907710106.

[10] V.N. Uversky, Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders, Front. Aging Neurosci. 7 (2015) 18, https://doi.org/10.3389/fnagi.2015.00018.

[11] S.F. Banani, H.O. Lee, A.A. Hyman, M.K. Rosen, Biomolecular condensates: organizers of cellular biochemistry, Nat. Rev. Mol. Cell Biol. 18 (2017) 285–298.

[12] S. Weber, C. Brangwynne, Inverse size scaling of the nucleolus by a concentration-dependent phase transition, Curr. Biol. 25 (5) (2015) 641–646, https://doi.org/10.1016/j.cub.2015.01.012.

[13] C.P. Brangwynne, C.R. Eckmann, D.S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, A.A. Hyman, Germline p granules are liquid droplets that localize by controlled dissolution/condensation, Science 324 (5935) (2009) 1729–1732, https://doi.org/10.1126/science.1172046.

[14] L.-P. Bergeron-Sandoval, N. Safaee, S. Michnick, Mechanisms and consequences of macromolecular phase separation, Cell 165 (5) (2016) 1067–1079, https://doi.org/10.1016/j.cell.2016.05.026.

[15] A.L. Darling, Y. Liu, C.J. Oldfield, V.N. Uversky, Intrinsically disordered proteome of human membrane-less organelles, Proteomics 18 (5–6) (2018) 1700193, https://doi.org/10.1002/pmic.201700193.

[16] V.N. Uversky, Protein intrinsic disorder-based liquid–liquid phase transitions in biological systems: complex coacervates and membrane-less organelles, Adv. Colloid Interf. Sci. 239 (2017) 97–114, https://doi.org/10.1016/j.cis.2016.05.012.

[17] V.N. Uversky, Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder, Curr. Opin. Struct. Biol. 44 (2017) 18–30, https://doi.org/10.1016/j.sbi.2016.10.015.

[18] S. Rauscher, R. Pomès, Molecular simulations of protein disorder, Biochem. Cell Biol. 88 (2) (2010) 269–290, https://doi.org/10.1139/O09-169.

[19] V.M. Burger, T. Gurry, C.M. Stultz, Intrinsically disordered proteins: where computation meets experiment, Polymers 6 (10) (2014) 2684–2719, https://doi.org/10.3390/polym6102684.

[20] R.B. Best, W. Zheng, J. Mittal, Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association, J. Chem. Theory Comput. 10 (11) (2014) 5113–5124, https://doi.org/10.1021/ct500569b.

[21] S. Piana, A.G. Donchev, P. Robustelli, D.E. Shaw, Water dispersion interactions strongly influence simulated structural properties of disordered protein states, J. Phys. Chem. B 119 (16) (2015) 5113–5123, https://doi.org/10.1021/jp508971m.

[22] J. Henriques, C. Cragnell, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment, J. Chem. Theory Comput. 11 (7) (2015) 3420–3431, https://doi.org/10.1021/ct501178z.

[23] S. Rauscher, V. Gapsys, M.J. Gajda, M. Zweckstetter, B.L. de Groot, H. Grubmüller, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment, J. Chem. Theory Comput. 11 (11) (2015) 5513–5524, https://doi.org/10.1021/acs.jctc.5b00736.

[24] J. Henriques, M. Skepö, Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the tip4p-d water model and the representativeness of protein disorder models, J. Chem. Theory Comput. 12 (7) (2016) 3407–3415, https://doi.org/10.1021/acs.jctc.6b00429.

[25] C. Cragnell, E. Rieloff, M. Skepö, Utilizing coarse-grained modeling and Monte Carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions, J. Mol. Biol. 430 (16) (2018) 2478–2492, https://doi.org/10.1016/j.jmb.2018.03.006.

[26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (1982) 105–132, https://doi.org/10.1016/0022-2836(82)90515-0.

[27] Y. Lin, S.L. Currie, M.K. Rosen, Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs, J. Biol. Chem. 292 (46) (2017) 19110–19120, https://doi.org/10.1074/jbc.M117.800466.

[28] C. Pak, M. Kosno, A. Holehouse, S. Padrick, A. Mittal, R. Ali, A. Yunus, D. Liu, R. Pappu, M. Rosen, Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein, Mol. Cell 63 (1) (2016) 72–85, https://doi.org/10.1016/j.molcel.2016.05.042.

[29] R. Williams, Z. Obradovic, V. Mathura, W. Braun, E. Garner, J. Young, S. Takayama, C. Brown, A. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, Pac. Symp. Biocomput. 2001 (6) (2001) 89–100.

[30] G.A. Naganagowda, T.L. Gururaja, M.J. Levine, Delineation of conformational preferences in human salivary statherin by 1H, 31P NMR and CD studies: sequential assignment and structure-function correlations, J. Biomol. Struct. Dyn. 16 (1) (1998) 91–107, https://doi.org/10.1080/07391102.1998.10508230.

[31] G.A. Elgavish, D.I. Hay, D.H. Schlesinger, 1H and 31P nuclear magnetic resonance studies of human salivary statherin, Int. J. Pept. Protein Res. 23 (3) (1984) 230–234, https://doi.org/10.1111/j.1399-3011.1984.tb02714.x.

[32] G. Goobes, R. Goobes, W.J. Shaw, J.M. Gibson, J.R. Long, V. Raghunathan, O. Schueler-Furman, J.M. Popham, D. Baker, C.T. Campbell, P.S. Stayton, G.P. Drobny, The structure, dynamics, and energetics of protein adsorption—lessons learned from adsorption of statherin to hydroxyapatite, Magn. Reson. Chem. 45 (S1) (2007) S32–S47, https://doi.org/10.1002/mrc.2123.

[33] N. Ramasubbu, L.M. Thomas, K.K. Bhandary, M.J. Levine, Structural characteristics of human salivary statherin: a model for boundary lubrication at the enamel surface, Crit. Rev. Oral Biol. Med. 4 (3) (1993) 363–370, https://doi.org/10.1177/10454411930040031501.

[34] P.A. Raj, M. Johnsson, M.J. Levine, G.H. Nancollas, Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization, J. Biol. Chem. 267 (9) (1992) 5968–5976.

[35] S.J. Whittington, B.W. Chellgren, V.M. Hermann, T.P. Creamer, Urea promotes polyproline II helix formation: implications for protein denatured states, Biochemistry 44 (16) (2005) 6269–6275, https://doi.org/10.1021/bi050124u.

[36] M. Kjaergaard, A.-B. Nørholm, R. Hendus-Altenburger, S.F. Pedersen, F.M. Poulsen, B.B. Kragelund, Temperature-dependent structural changes in intrinsically disordered proteins: formation of α-helices or loss of polyproline II? Protein Sci. 19 (8) (2010) 1555–1564, https://doi.org/10.1002/pro.435.

[37] M. Abu-Hamdiyyah, The effect of urea on the structure of water and hydrophobic bonding, J. Phys. Chem. 69 (8) (1965) 2720–2725, https://doi.org/10.1021/j100892a039.

[38] W. Bruning, A. Holtzer, The effect of urea on hydrophobic bonds: the critical micelle concentration of *n*-dodecyltrimethylammonium bromide in aqueous solutions of urea1, J. Am. Chem. Soc. 83 (23) (1961) 4865–4866, https://doi.org/10.1021/ja01484a044.

[39] U. Thapa, K. Ismail, Urea effect on aggregation and adsorption of sodium dioctylsulfosuccinate in water, J. Colloid Interface Sci. 406 (2013) 172–177, https://doi.org/10.1016/j.jcis.2013.06.009.

[40] J. Broecker, S. Keller, Impact of urea on detergent micelle properties, Langmuir 29 (27) (2013) 8502–8510, https://doi.org/10.1021/la4013747.

[41] A. Malliaris, J. Le Moigne, J. Sturm, R. Zana, Temperature dependence of the micelle aggregation number and rate of intramicellar excimer formation in aqueous surfactant solutions, J. Phys. Chem. 89 (12) (1985) 2709–2713, https://doi.org/10.1021/j100258a054.

[42] R. Zana, C. Weill, Effect of temperature on the aggregation behaviour of nonionic surfactants in aqueous solutions, J. Phys. Lett. 46 (20) (1985) 953–960.

[43] C. Moitzi, I. Portnaya, O. Glatter, O. Ramon, D. Danino, Effect of temperature on self-assembly of bovine β-casein above and below isoelectric pH. Structural analysis by cryogenic-transmission electron microscopy and small-angle x-ray scattering, Langmuir 24 (7) (2008) 3020–3029, https://doi.org/10.1021/la702802a.

[44] M. Kenward, M.D. Whitmore, A systematic Monte Carlo study of self-assembling amphiphiles in solution, J. Chem. Phys. 116 (8) (2002) 3455–3470, https://doi.org/10.1063/1.1445114.

[45] D. Franke, M.V. Petoukhov, P.V. Konarev, A. Panjkovich, A. Tuukkanen, H.D.T. Mertens, A.G. Kikhney, N.R. Hajizadeh, J.M. Franklin, C.M. Jeffries, D.I. Svergun, *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions, J. Appl. Crystallogr. 50 (4) (2017) 1212–1225, https://doi.org/10.1107/S1600576717007786.

[46] D. Orthaber, A. Bergmann, O. Glatter, SAXS experiments on absolute scale with Kratky systems using water as a secondary standard, J. Appl. Crystallogr. 33 (2) (2000) 218–225, https://doi.org/10.1107/S0021889899015216.

[47] J. Reščič, P. Linse, MOLSIM: a modular molecular simulation software, J. Comput. Chem. 36 (16) (2015) 1259–1274, https://doi.org/10.1002/jcc.23919.

[48] M. Lal, Monte Carlo computer simulation of chain molecules. I, Mol. Phys. 17 (1) (1969) 57–64.

[49] N. Madras, A.D. Sokal, The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk, J. Stat. Phys. 50 (1) (1988) 109–186, https://doi.org/10.1007/BF01022990.

[50] F.T. Wall, F. Mandel, Macromolecular dimensions obtained by an efficient Monte-Carlo method without sample attrition, J. Chem. Phys. 63 (11) (1975) 4592–4595.

Supplemental information for Assessing the intricate balance of intermolecular interactions upon self-association of intrinsically disordered proteins

Ellen Rieloff[a,*], Mark Tully[b], Marie Skepö[a,*]

[a] *Theoretical Chemistry, Lund University, POB 124, SE-221 00 Lund, Sweden*
[b] *European Synchrotron Radiation Facility (ESRF), Grenoble, France*

Analysis of Small-angle X-ray scattering data

Here we present collected SAXS curves and additional information regarding the determination of forward scattering and radius of gyration for the data collected at 20 °C with 10 and 150 mM NaCl. The data at other salt concentrations and temperatures were treated in the same way. Figure S1 shows the scattering curves for Statherin with increasing protein concentration measured at 20 °C, for 150 and 10 mM NaCl. At higher concentrations than presented in the figure, a clear depression at low q was shown, and therefore such data was excluded from analysis. The forward scattering and radius of gyration were determined by both the Guinier method and from the pair distance distribution function, P(r). Guinier plots with fits to the used range are presented in Figure S2 for the data at 150 mM NaCl and in Figure S3 for the data at 10 mM NaCl. The used range in the Guinier method was limited to $qR_\mathrm{g} < 0.8$, or extended to $qR_\mathrm{g} < 1.0$ when appropriate, since that is usually the linear region for an IDP [1]. The figures also include the fits in the P(r) analysis. The resulting values are presented in Table S1 and Table S2. Overall the agreement between the two methods are good, although the radius of gyration from the pair distance distribution is slightly larger. Since it is known that the Guinier law is less appropriate for describing an unfolded chain and therefore can underestimate the size of intrinsically disordered proteins, we have presented the values from the pair distribution function in the article.

The molecular weight, M_w, was calculated using the following equation:

$$M_\mathrm{w} = \frac{I_0 \cdot I_{0\mathrm{w,ref}} \cdot N_\mathrm{A}}{I_{0\mathrm{w,meas}} \cdot c([\rho_\mathrm{p} - \rho_\mathrm{s}]\nu_\mathrm{p})} \tag{1}$$

where the forward scattering I_0 is given in arbitrary units, $I_{0\mathrm{w,ref}}$ is the absolute scattering of water, N_A is the Avogadro constant, $I_{0\mathrm{w,meas}}$ the measured scattering of water in arbitrary units, and c the protein

[*] Corresponding author
Email addresses: `ellen.rieloff@teokem.lu.se` (Ellen Rieloff), `marie.skepo@teokem.lu.se` (Marie Skepö)

concentration. The electron density of the protein, ρ_p, was determined from the number of electrons in the protein and the molecular weight, while the electron density of the solvent, ρ_s, was calculated with MulCh [2] based on the Tris and NaCl concentrations. The partial specific volume of the protein, ν_p, was calculated from the amino acid sequence using Sednterp [3], assuming no effect from phosphorylations.

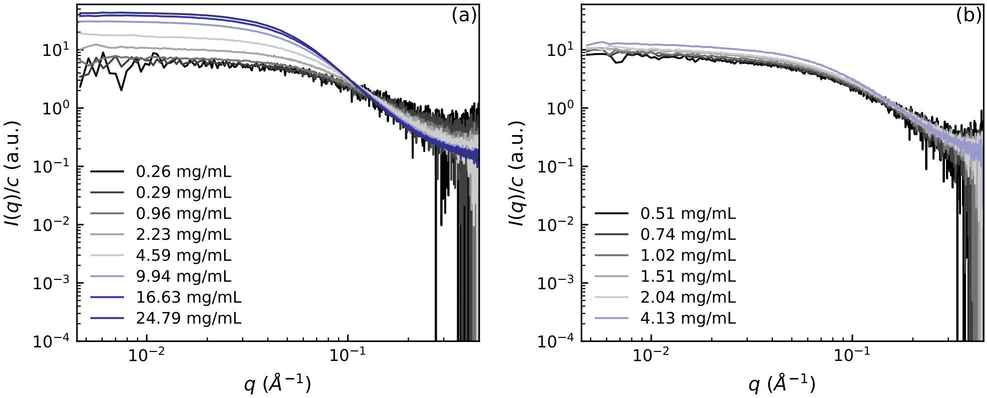


Figure S1: Overlayed scattering curves for Statherin with (a) 150 mM NaCl and (b) 10 mM NaCl, and 20 mM Tris, pH 8.1, at 20 °C.

Table S1: Forward scattering, I_0, and radius of gyration, R_g, determined both by the Guinier approximation and from the pair distribution function, for the data at 150 mM NaCl and 20 °C.

c (mg/mL)	$I_{0,\mathrm{Guinier}}/c$ (a.u.)	$I_{0,\mathrm{P(r)}}/c$ (a.u.)	$R_{\mathrm{g,Guinier}}$ (Å)	$R_{\mathrm{g,P(r)}}$ (Å)
0.26	5.9 ± 0.1	6.0 ± 0.1	17.1 ± 0.6	19.0 ± 0.4
0.29	6.5 ± 0.1	6.4 ± 0.1	20.7 ± 0.9	20.1 ± 0.3
0.96	7.3 ± 0.1	7.4 ± 0.1	20.0 ± 0.2	20.8 ± 0.2
2.23	10.5 ± 0.1	10.5 ± 0.1	22.5 ± 0.2	23.1 ± 0.2
4.59	17.0 ± 0.1	17.1 ± 0.1	25.8 ± 0.2	26.9 ± 0.3
9.94	30.6 ± 0.1	30.7 ± 0.1	31.4 ± 0.8	31.8 ± 0.1
16.63	39.5 ± 0.1	39.7 ± 0.1	32.2 ± 0.3	32.7 ± 0.1
24.79	44.4 ± 0.1	45.4 ± 0.1	31.9 ± 0.6	33.2 ± 0.1

Table S2: Forward scattering, I_0, and radius of gyration, R_g, determined both by the Guinier approximation and from the pair distribution function, for Statherin at 10 mM NaCl and 20 °C.

c (mg/mL)	$I_{0,\mathrm{Guinier}}/c$ (a.u.)	$I_{0,\mathrm{P(r)}}/c$ (a.u.)	$R_{\mathrm{g,Guinier}}$ (Å)	$R_{\mathrm{g,P(r)}}$ (Å)
0.51	6.7 ± 0.1	6.8 ± 0.1	19.8 ± 0.9	21.9 ± 0.8
0.74	7.5 ± 0.1	7.5 ± 0.1	22.7 ± 0.5	24.2 ± 0.7
1.02	8.0 ± 0.1	8.0 ± 0.1	22.0 ± 0.3	23.1 ± 0.4
1.51	8.8 ± 0.1	9.0 ± 0.1	22.2 ± 0.3	23.9 ± 0.3
2.04	9.4 ± 0.1	9.5 ± 0.1	21.9 ± 0.3	23.4 ± 0.3
4.13	11.3 ± 0.1	11.5 ± 0.1	21.8 ± 0.2	23.1 ± 0.1

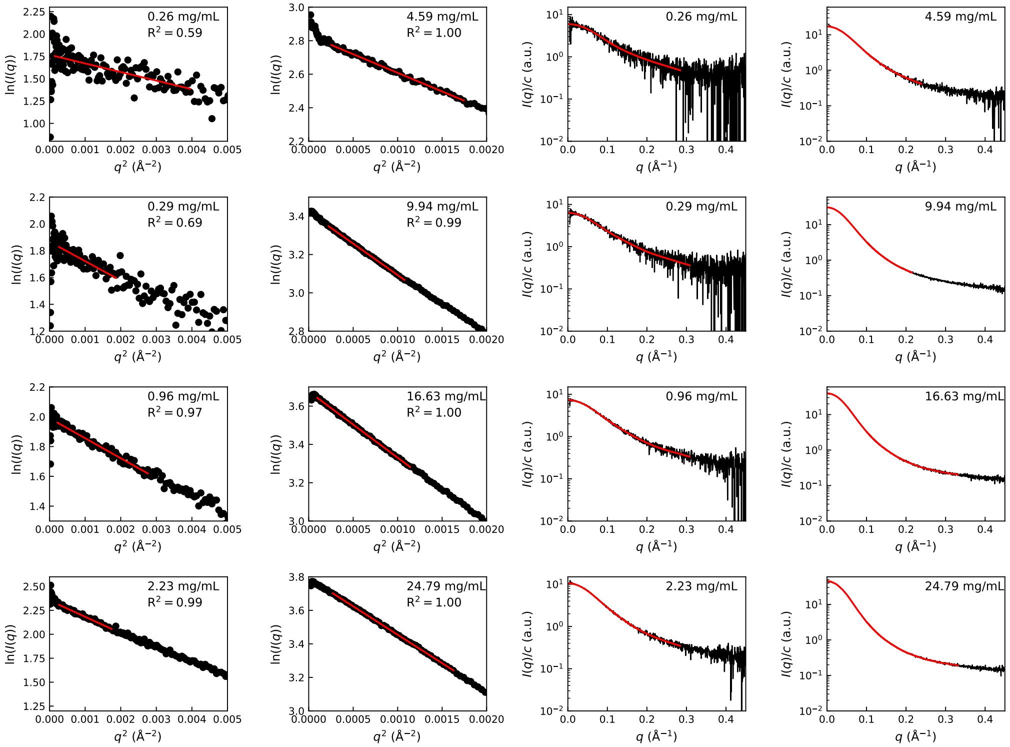


Figure S2: Guinier plots (the two left columns) and SAXS curves with the fits obtained in the P(r) analysis (the two right columns) for the reference system, obtained with 150 mM NaCl, 20 mM Tris, pH 8.1, at 20 °C. The red straight lines in the Guinier plots are the Guinier fits in the used range. The red curves are obtained in the indirect transform for obtaining P(r), using the ATSAS package [4].

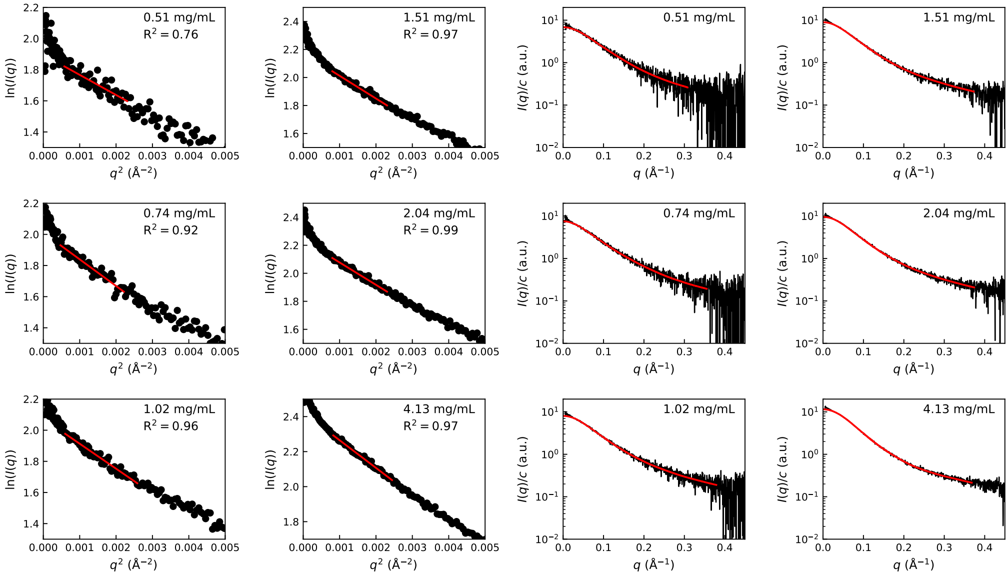


Figure S3: Guinier plots with red lines corresponding to the Guinier approximation in the used range (the two left columns) and SAXS curves with the fits obtained in the P(r) analysis given in red (the two right columns) for Statherin with 10 mM NaCl, 20 mM Tris, pH 8.1, at 20 °C. The red straight lines in the Guinier plots are the Guinier fits in the used range. The red curves are obtained in the indirect transform for obtaining P(r), using the ATSAS package [4].

Circular Dichroism data

To provide an estimate of the variation in the circular dichroism data, Figure S4 shows how the smoothened data achieved by applying a Savitzky–Golay filter relates to the raw data for two replicates at 4 and 28 °C. For each replicate a new sample was prepared and the measurements of the different replicates were made on different days. At 4 °C the agreement between the two replicates is excellent, while there is a small difference between the replicates at 28 °C. Factors contributing to the variation involves noise as well as uncertainties in the measured concentration.

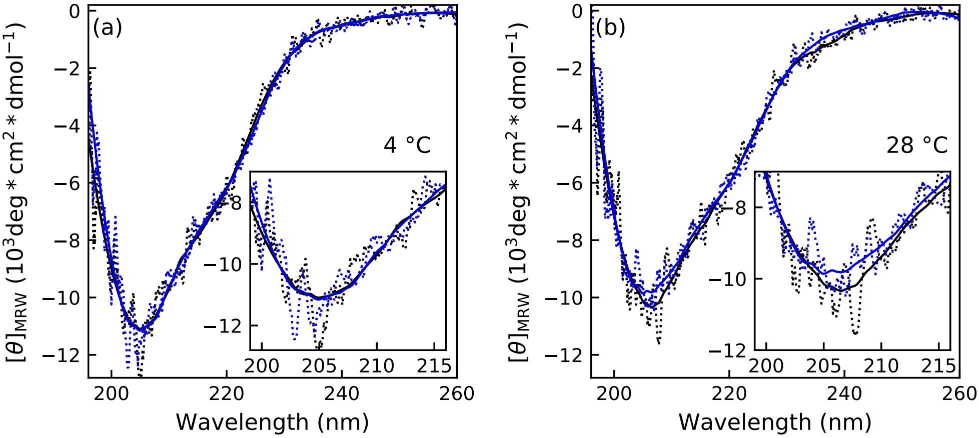


Figure S4: Raw data (dotted lines) and smoothened data (solid lines) from two different circular dichroism measurements (blue and black) for Statherin at (a) 4 °C and (b) 28 °C, in 20 mM phosphate buffer, 150 mM NaF, pH 8. The insets are enlargements of the data around the minimum.

References

[1] V. Receveur-Brechot, D. Durand, How random are intrinsically disordered proteins? a small angle scattering perspective, Curr. Protein Pept. Sci. 13 (1) (2012) 55–75. doi:doi:10.2174/138920312799277901.

[2] A. E. Whitten, S. Cai, J. Trewhella, *MULCh*: modules for the analysis of small-angle neutron contrast variation data from biomolecular assemblies, J. Appl. Crystallogr. 41 (1) (2008) 222–226. doi:10.1107/S0021889807055136.

[3] T. Hurton, A. Wright, G. Deubler, B. Bashir, Sedimentation interpretation program, 20120828 BETA, based on the original program by D. B. Hayes and T. Laue and J. Philo. Available at http://rasmb.org/sednterp/.

[4] D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, D. I. Svergun, *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions, J. Appl. Crystallogr. 50 (4) (2017) 1212–1225. doi:10.1107/S1600576717007786.

Paper III

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

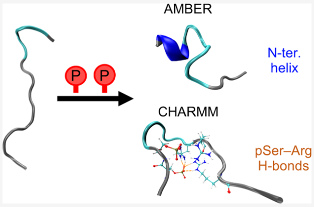
Phosphorylation of a Disordered Peptide—Structural Effects and Force Field Inconsistencies

Ellen Rieloff* and Marie Skepö*

Cite This: J. Chem. Theory Comput. 2020, 16, 1924−1935 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Phosphorylation is one of the most abundant types of post-translational modifications of intrinsically disordered proteins (IDPs). This study examines the conformational changes in the 15-residue-long N-terminal fragment of the IDP statherin upon phosphorylation, using computer simulations with two different force fields: AMBER ff99SB-ILDN and CHARMM36m. The results from the simulations are compared with experimental small-angle X-ray scattering (SAXS) and circular dichroism data. In the unphosphorylated state, the two force fields are in excellent agreement regarding global structural properties such as size and shape. However, they exhibit some differences in the extent and type of the secondary structure. In the phosphorylated state, neither of the force fields performs well compared to the experimental data. Both force fields show a compaction of the peptide upon phosphorylation, greater than what is seen in SAXS experiments, although they differ in the local structure. While the CHARMM force field increases the fraction of bends in the peptide as a response to strong interactions between the phosphorylated residues and arginines, the AMBER force field shows an increase of the helical content in the N-terminal part of the peptide, where the phosphorylated residues reside, in better agreement with circular dichroism results.



1. INTRODUCTION

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure in solution under physiological conditions.[1,2] Despite this, they are functional and participate in the regulation of many biological processes,[3,4] in which disorder can enable interactions of high specificity coupled with low affinity.[5] These interactions are in part regulated by post-translational modifications, such as phosphorylation, which is reversible. Phosphorylation sites are prevalent in both disordered regions and IDPs.[5−7]

Disordered phosphoproteins regulate, for example, physiological biomineralization, by involvement at various stages.[8] Statherin, a saliva IDP, is involved in the regulation of tooth mineralization by inhibiting spontaneous precipitation and crystal growth of calcium phosphates.[9−11] Caseins and osteopontin are examples of IDPs that can sequester amorphous calcium phosphate through interaction with phosphorylated residues, and by this, stabilize supersaturated fluids.[12−14] IDPs susceptible to phosphorylation can also be involved in diseases, for instance the tau protein, for which abnormal hyperphosphorylation has been related to amyloid fibril formation in Alzheimer's disease.[15]

The addition of a phosphoryl group changes the properties of the residue, the most prominent change being the addition of a double-negative charge at physiological pH. The phosphoryl group also allows for multiple hydrogen bonds, which can drastically affect the protein conformation or interaction with a binding partner, hence affecting the affinity.[16] The possible effects of phosphorylation involve transition between disorder and order, changes in association state, and activation or inhibition of a protein.[16]

The most occurring phosphorylated residue is phosphoserine,[17] and it is known to act as either a stabilizer or a destabilizer of α-helices, depending on the position in the helix, and residues in the surroundings.[18,19] In the N-terminal end of a helix, phosphoserine acts as a stabilizer because of hydrogen-bonding with the backbone NH groups that do not take part in the i, $i + 4$ hydrogen bonding pattern characteristic of α-helices, and electrostatic interaction with the helix macrodipole.[18] The presence of a phosphoserine four steps away from a lysine also stabilizes helices, through formation of a salt bridge between the phosphate group and the positively charged side chain of lysine. Other positively charged side chains are suggested to have the same type of stabilizing effect.[19] Phosphoserines have also been shown to be involved in strong interactions with arginines, through salt bridge formation with the guanidinium group of the side chain.[20,21] These interactions can play an important role in the conformational response and recognition.[16]

Since IDPs possesses vast conformational ensembles, their structure can be rather challenging to study experimentally. Hence, computer simulations have emerged as a useful tool to complement experiments.[22,23] During recent years, there has been considerable advancements in atomistic simulations of

Received: November 28, 2019
Published: February 12, 2020

IDPs, because of the development and justification of force fields and water models against experimental results.[24−28] Therefore, atomistic simulation studies of IDPs have become more common and also phosphorylated IDPs have been studied with various force fields.[29−38] However, the parameters for phosphorylated residues precede many of the more recent optimizations of force fields for IDPs, and might therefore not work as intended with the latest force field developments. Hence, there is a need to examine the performance of newer force fields with the extensions available for phosphorylated residues. To the authors' knowledge, only few studies have assessed the performance of force fields for phosphorylated residues. Recently Vymětal, Jurásková and Vondrášek presented a study of the effects of phosphorylation on dipeptides, showing inconsistencies in the conformational changes among the tested force fields:[39] AMBER ff99SB[40] extended by the phosaa10 parameters for phosphorylated residues, developed by Homeyer et al.[41] and Steinbrecher et al.,[42] AMBER ff03[43] with phosphorylated parameters from Forcefield_PTM,[44] and CHARMM36m.[45] Although the intrinsic conformational preferences of phosphorylated residues are of importance for the conformational changes in a protein, long-range interactions with other residues, such as salt bridge formation, can play a major role. Therefore, it is necessary to systematically investigate force field effects in longer peptide sequences.

In this study, the effect of phosphorylated serines in a model peptide is investigated using two different force fields: (i) AMBER ff99SB-ILDN[46] with the TIP4P-D[25] water model, extended by the phosaa10 parameters, and (ii) CHARMM36m[45] with the CHARMM-modified TIP3P water model,[47] which already contains parameters for phosphorylated residues. The model peptide used in this study is the 15-residue-long N-terminal fragment of the saliva IDP statherin. Previous studies on this fragment have shown that phosphorylation affects the secondary structure, and that the unphosphorylated peptide has a reduced ability to adhere to hydroxyapatite surfaces and to inhibit mineralization.[11] Hence, phosphorylation regulates the functionality of statherin, and it is therefore of interest to further investigate the possible conformational effects induced by phosphorylation. The results from the simulations are compared with experimental data collected by small angle X-ray scattering (SAXS) and circular dichroism (CD), to assess the performance of the force fields regarding both overall shape and secondary structure.

2. METHODS

2.1. Computational Methods. The initial configuration of the nonphosphorylated peptide (SN15n) was built as a linear chain in PyMOL,[48] whereas the phosphorylated peptide (SN15p) was built as a linear chain in Avogadro 1.2.0,[49] in which the structure was optimized using the auto-optimization tool. The molecular dynamics simulations were performed using the GROMACS package version 2018.4,[50−54] with two different force fields and water models: (i) AMBER ff99SB-ILDN[46] with the TIP4P-D[25] water model and parameters for the phosphorylated residues from Homeyer et al.[41] and Steinbrecher et al.,[42] as presented in the parameter set phosaa10 found in, for example, the Supporting Information to Steinbrecher et al.,[42] and (ii) CHARMM36m[45] with the CHARMM-modified TIP3P water model,[47] using the included parameters for phosphorylated residues.

The peptide was solvated in a rhombic dodecahedron box, having a minimum distance between the peptide and the box edges of 10 Å. One chloride ion or three sodium ions were added to neutralize the system, for SN15n and SN15p, respectively. The number of solvent molecules is specified in Table 1.

Table 1. System Specification[a]

peptide	force field[b]	N_{water}	simulation length (μs)
SN15n	A99	8839	2.0 + 3.0 + 3.4 + 2.0 + 4.0
SN15n	C36	8861	3.0 + 3.0 + 4.4 + 4.0 + 3.0
SN15p	A99	9703	4.4 + 4.4 + 4.4 + 4.4 + 4.4
SN15p	C36	9508	4.0 + 4.0 + 4.0 + 4.0 + 4.0

[a]Number of water molecules and the simulation length of each replicate. [b]A99 = AMBER ff99SB-ILDN with the TIP4P-D water model, C36 = CHARMM36m with the CHARMM-modified TIP3P water model.

Periodic boundary conditions were employed in all directions. The Verlet leapfrog algorithm[55] with a time step of 2 fs was used to integrate the equations of motion. Nonbonded interactions were treated with a Verlet list cutoff scheme. The short-ranged interactions were calculated using neighbor lists with cutoffs of 10 and 12 Å, for AMBER and CHARMM force fields, respectively. When using the CHARMM force field, the Lennard-Jones interactions were switched off smoothly (force-switch) between 10 and 12 Å. Long-ranged dispersion corrections were applied to energy and pressure when using the AMBER force field. Long-ranged electrostatic interactions were treated by particle mesh Ewald[56] with a cubic interpolation and 1.6 Å grid spacing. Solute and solvent were separately coupled to temperature baths at 298 K using the velocity rescaling thermostat[57] with a 0.1 ps relaxation time. The pressure was set to 1 bar by the Parrinello−Rahman pressure coupling[58] with a 2 ps relaxation time and 4.5×10^{-5} bar^{-1} isothermal compressibility. Using the LINCS algorithm,[59] the bond lengths were constrained for all bonds in the AMBER force field simulations, and only for bonds with hydrogen atoms in the CHARMM simulations.

Energy minimization was performed by the steepest descent algorithm until the system was converged within the available machine precision. Initiation of replicates was performed separately in two steps using position restraints on the peptide. The first step was 500 ps of *NVT* simulation (constant number of particles, volume, and temperature) performed to stabilize the temperature, followed by the second step of 1000 ps of *NPT* simulation (constant number of particles, pressure, and temperature) to stabilize the pressure. The production run comprised five replicates of at least 2 μs each in the *NPT* ensemble. Exact simulation times used are presented in Table 1. Energies and coordinates were saved every 10 ps.

2.2. Analysis. Simulation analyses were performed using GROMACS package version 2018.4,[50−54] the MDTraj Python library version 1.9.3,[60] and the DSSP program version 2.2.1.[61] Error estimates of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) were calculated using block averaging analysis as implemented in the *gmx analyze* routine in GROMACS. SAXS intensities were calculated using CRYSOL version 2.8.3.[62] The energy landscapes were calculated using principal component analysis following the Campos and Baptista approach,[63] with the differences described by Henriques et al.[26] Representative snapshots from the simulations were produced using VMD 1.9.3.[64−66]

Convergence was checked by performing each simulation in five replicates and comparing their end-to-end distance and radius of gyration distributions, and their energy landscapes, as well as observing the auto-correlation function and block average error estimate of the concatenated simulation. The reader is referred to the Supporting Information for a detailed assessment of the convergence and the sampling quality. Before final analysis, the first 0.15 μs were removed from each replicate of the phosphorylated peptide, as this time span showed signs of equilibration. For the nonphosphorylated peptide, the initial time was not removed, as the equilibration was fast enough for the effect to be negligible.

2.3. Experimental Methods. *2.3.1. Sample Preparation.* 20 mM Tris [Ultrapure >99.9%, CAS 77-86-1; Saveen Werner AB] buffer at pH 7.5 (20 °C) with 150 mM NaCl [AnalaR NORMAPUR, CAS 7647-14-15; VWR Chemicals, Belgium] was prepared with Milli-Q water, and the pH was adjusted by drop-wise addition of 1 M HCl. This buffer was used for the SAXS samples. As Tris absorbs light at low wavelengths, 20 mM phosphate buffer ($Na_2HPO_4{\cdot}2H_2O$ [Reag. Ph. Eur., CAS 10028-24-7; Sigma-Aldrich, Germany] and $NaH_2PO_4{\cdot}2H_2O$ [ACS reagent, CAS 10049-21-5; Sigma-Aldrich, Germany]), pH 7.5, prepared with Milli-Q water and drop-wise addition of 1 M NaOH to adjust the pH, was used for the CD samples.

Peptide in the form of a lyophilized powder (synthesized by TAG Copenhagen A/S, Denmark) was dissolved in buffer, and dialyzed using membranes with a cutoff of 500−1000 Da (Spectra/Por, Biotech-Grade CE Dialysis Tubing) in buffer of a volume ≥ 400 times the sample volume while stirring. The buffer was exchanged four times during a total dialysis time of 48 h, in which the first 7 h of dialysis was performed in room temperature, and the remaining time at 6 °C.

For the CD samples, a concentrated stock peptide solution was purified and then diluted to the desired concentration for measurements using 20 mM phosphate buffer with 150 mM NaF [99.5%, CAS 7681-49-4; VWR Chemicals, Germany], as also chloride ions absorb strongly below 200 nm.

2.3.2. Small-Angle X-ray Scattering. The SAXS experiments were performed at beamline B21 of the Diamond Light Source, United Kingdom. The stock solution and dialysis buffer were centrifuged at 14,000 rpm at 8 °C for 4 h to remove potential large aggregates and/or impurities, before diluting to the desired concentrations (a series of approximately 1, 2, 4, and 6 mg/mL). Because of the limitations of the NanoDrop 1000 instrument, the protein concentration was determined by absorption at 257 nm, using an extinction coefficient of 390 cm^{-1} M^{-1}. This is based on the peptide containing two phenylalanines and that the absorption of phenylalanine at 257 nm is 195 cm^{-1} M^{-1}.[67] The concentration determined at 257 nm was approximately equal to the concentration determined at 214 nm. Because of low absorbance yielding unreliable results, for the SN15n peptide, the absorption was only measured for the highest concentration, and the other concentrations were calculated based on the dilution scheme.

The distance between the sample and the Pilatus 4M detector was 4.014 m at 12.4 keV, corresponding to a q-range of 0.0034−0.44 Å. The scattering vector, q, is defined as $q = 4\pi \sin \theta/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident beam, 1 Å. The measurements were performed using the BioSAXS sample robot, loading the samples into a flow-through quartz capillary. Fifteen consecutive frames were recorded using an exposure time of 2 s each, at 20 °C. The dialysis buffer, that is the background, was measured first in each concentration series. For the lowest concentrations, measurements were performed in several replicates, and then final averages were determined in the data processing stage.

Data processing and analysis were performed using the ATSAS package.[68] Prior to averaging and buffer subtraction, the consecutive frames were checked for signs of radiation damage, and affected frames were removed. The forward scattering, I_0, and radius of gyration, R_g, were determined both from Guinier analysis and the pair distance distribution, $P(r)$. From I_0, the molecular weight was calculated using a conversion factor determined from a measurement on a bovine serum albumin standard.

2.3.3. Circular Dichroism. Before measurement, the sample was filtered through a 0.22 μm filter (Millex-GV, Merck Millipore Ltd.), and the peptide concentration was determined from the absorption at 214 nm measured with a NanoDrop 2000, using an extinction coefficient of 24,000 M^{-1} cm^{-1}.[69] The protein concentration was ∼0.17 mg/mL. CD spectra were recorded between 185 and 260 nm for the samples and the buffer at 20 °C using a JASCO J-715 instrument with a PTC-348WI Peltier type cell holder for temperature control, in a 0.1 mm quartz cuvette (Hellma Analytics). The scanning speed was 20 nm/min, the response time 2 s, the bandwidth 1.0 nm, the data pitch 0.1 nm, and the sensitivity 100 mdeg. Each of the spectra was averaged over five recordings.

The reported ellipticity is expressed as mean residue ellipticity, defined as

$$[\theta]_{\mathrm{MRW}} = \theta{\cdot}\mathrm{MRW}/(10{\cdot}d{\cdot}c) \tag{1}$$

where θ is the observed ellipticity (mdeg), d the path length of the cell (cm), and c the protein concentration (mg/mL). The mean residue weight, MRW, is the molecular weight (Da) divided by the number of peptide bonds.

To assess the partition of secondary structural elements in the peptides, the data were analyzed with BeStSel[70,71] through a web-server (http://bestsel.elte.hu/index.php).

3. RESULTS AND DISCUSSION

An N-terminal fragment of statherin, namely, the first 15 amino acids with the sequence presented in Figure 1, has been studied in the unphosphorylated state (SN15n), and with two phosphorylated serine residues (SN15p), using two different force fields: AMBER ff99SB-ILDN (hereafter A99) and CHARMM36m (hereafter C36). The results for the two different peptides are presented together in the same figures, but are initially discussed separately. First, the force fields are compared for SN15n and SN15p separately, followed by a discussion regarding the effect of phosphorylation in the SN15 sequence, as well as a comparison to experimental data.

SN15n: D-S-S-E-E-K-F-L-R-R-I-G-R-F-G
SN15p: D-pS-pS-E-E-K-F-L-R-R-I-G-R-F-G

Figure 1. Sequences of the two peptides SN15n and SN15p. Negatively charged amino acids are marked in red, and positively charged residues in blue. Note that each phosphorylated serine has a charge of −2e.

3.1. SN15n: The Nonphosphorylated Peptide. Both the end-to-end distance and the radius of gyration, shown in Figure 2, sample rather broad distributions for the SN15n peptide. This is the expected behavior for IDPs, which generally exhibit a wide range of conformations. For both properties, there is good agreement between the two different force fields, suggesting that they sample the same global properties. The calculated SAXS intensities, presented in Figure 3 as a dimensionless Kratky plot,

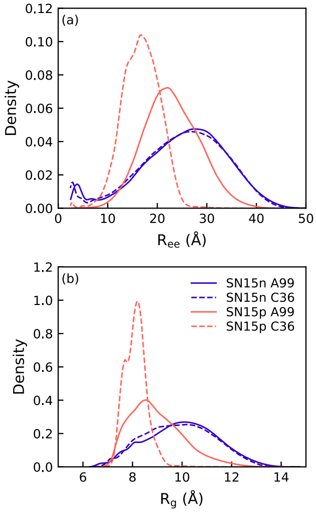


Figure 2. Density estimate of the end-to-end distance, R_{ee} (a), and the radius of gyration, R_g (b), of SN15n and SN15p simulated with AMBER ff99SB-ILDN and CHARMM36m, obtained using a Gaussian kernel estimator. The legend applies to both panels.

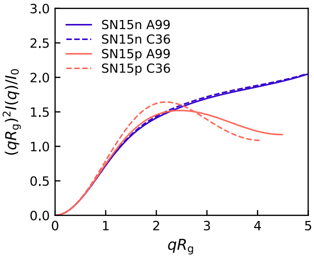


Figure 3. Dimensionless Kratky plot of SN15n and SN15p simulated with AMBER ff99SB-ILDN and CHARMM36m.

are also indistinguishable, which is expected when the global properties are the same. Furthermore, the Kratky plot shows that the behavior of SN15n resembles that of a random coil, which is typical for many IDPs.

Regarding the local properties, more specifically the secondary structure, disagreement between the two force fields is visible. It has already been pointed out, for example by Zerze et al.[72] when comparing three generations of AMBER03 force fields, that agreement on global properties does not imply agreement on local properties. In this case, the secondary structure determined by DSSP, presented in Figure 4, shows that for both force fields, the peptide is dominated by an irregular structure, such that the structure content overall is rather low. However, the A99 force field gives a larger helical content, specifically more 3_{10}-helix, between residues 3 and 8 than C36. Another view of the secondary structure is provided by the Ramachandran plots shown in Figure 5, where population at ϕ = −75 ± 20°, ψ = 145 ± 20° signalizes polyproline type II (PPII) structure, which is not sampled by the DSSP program. The Ramachandran plot shows a larger distribution in accessible ϕ and ψ values for A99 than C36. In addition, for A99 there are two clear maxima in the areas usually corresponding to β-strand and PPII structure, and a third smaller maximum in the helical region, indicating a 3_{10}-helix. The C36 simulation shows only one distinct maximum, in the PPII area, and a smaller maximum corresponding to α-helical structure.

Qualitatively, the secondary structure analysis from DSSP and the Ramachandran plot agree rather well, besides the Ramachandran plot capturing the PPII structure. The A99 simulation shows more sampling in the β-strand region of the Ramachandran plot than the C36 simulation, and slightly more β-sheet and β-bridge in the DSSP analysis. The identification of helicity is also in good agreement between the two methods, especially as A99 contains more 3_{10}-helical than α-helical structure, whereas the opposite is true for C36, a conclusion from both the Ramachandran plots and the DSSP analysis.

The rather unstructured peptide conformation is further confirmed by the contact map in Figure 6. Some residues have a higher probability of being close to each other than to others, but overall it is in good agreement with a broad and interchangeable conformational ensemble without clear specificity. There are some differences between the two force fields, the most apparent one being a 40% probability of having Arg10 and Phe14 close (smallest distance between atoms <4 Å) in A99, whereas the corresponding probability of C36 is 25%. This close distance is probably related to the Gly12 often being in a bend, as shown by Figure 4. In C36, a bend centered around Phe7 is associated with an increased probability of having residues Glu5 and Arg9 close together, an interaction that is electrostatically favorable. For A99, in the segment of residues 2−9, there is an increased probability of being close to residues three neighbors away, which is connected to the higher occurrence of 3_{10}-helix in this segment.

Although A99 showed a larger helical content in the N-terminal part of the peptide, this difference is not visible in the number of intrapeptide hydrogen bonds, shown in Figure 7. The distribution of the number of hydrogen bonds in a conformation is very similar between the two force fields, especially focusing on the type of hydrogen bonds that characterize helices.

To summarize the similarities and differences between the two force fields, energy landscapes complemented with representative structures of each minimum, shown in Figure 8, provide a good overview of the simulated system. It is worth pointing out that the first two components only account for approximately 40−50% of the variance in the simulations. Hence this analysis does not provide a complete picture of all the conformational classes. Despite that, it still provides an overview and is adequate for a brief comparison between the two force fields. From the figure, it is clear that the conformational space is similar for the two force fields, and that the most common conformation is rather stretched and irregular. Whereas the percentage stating the part of all sampled conformations belonging to each basin in the energy landscape is approximate, it remains clear that the conformations with more secondary structure are found in much less-populated minima. Whereas both force fields show a small share of the β-sheet structure, only A99 show conformations with a large part of the peptide in a 3_{10}-helix. This is in agreement with the DSSP analysis.

3.2. SN15p: The Phosphorylated Peptide. In the case of the phosphorylated peptide, SN15p, the differences between the two force fields are much larger than for the nonphosphorylated counterpart. Figure 2 shows that the probability distributions of both R_{ee} and R_g are more narrow and centered around smaller

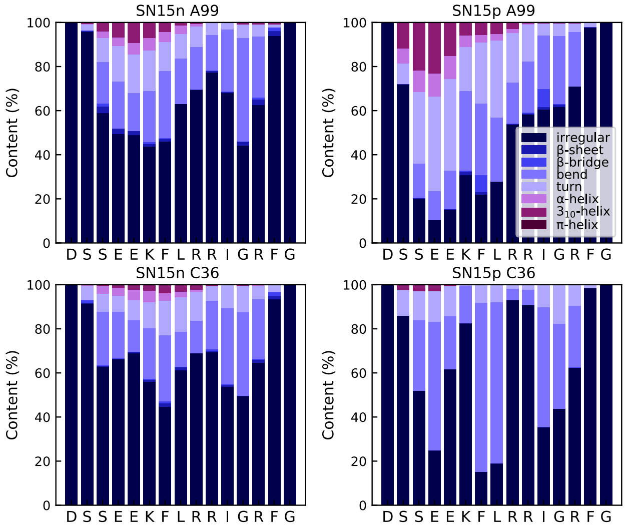


Figure 4. Stacked bar chart of the secondary structure content determined by DSSP of each amino acid in the SN15n (left column) and SN15p (right column) peptides, simulated with AMBER ff99SB-ILDN (top row) and CHARMM36m (bottom row). The legend applies to all panels.

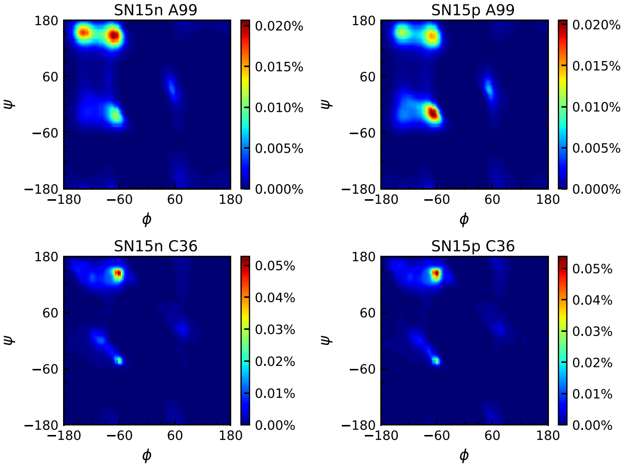


Figure 5. Ramachandran plots of SN15n (left column) and SN15p (right column) simulated with AMBER ff99SB-ILDN (top row) and CHARMM36m (bottom row). The color scale shows the population density.

values in the case of C36. The Kratky plot (Figure 3) also indicates a more compact, globule-like structure for C36 than for A99.

The differences on the global structural properties continues on the local properties. From Figure 4, it is clear that the A99 force field gives a much larger helical content in the N-terminal end (between residues two and nine) than C36. Instead, C36 shows a larger content of bends, especially in the middle of the peptide. Overall, A99 gives a wider range of different structures, as also more β-strands are sampled. This is also evident from the Ramachandran plot in Figure 5, where A99 shows a larger distribution than C36, in accordance with the nonphosphorylated peptide. For A99, the largest maximum is located in the helical region and indicates more 3_{10}-helix than α-helix, in agreement with the DSSP analysis. Other maxima are located in the β-strand and PPII region. As for SN15n, C36 shows a strong maximum in the PPII region and a secondary maximum in the α-helical region.

Both force fields show some specific contacts involving the phosphorylated serines, although the effect is much stronger for C36. Figure 6 reveals that around 85% of the sampled conformations using C36 have a distance < 4 Å between atoms in residue pSer2 and Arg13. A closer investigation shows that hydrogen bonds are formed between the side group of

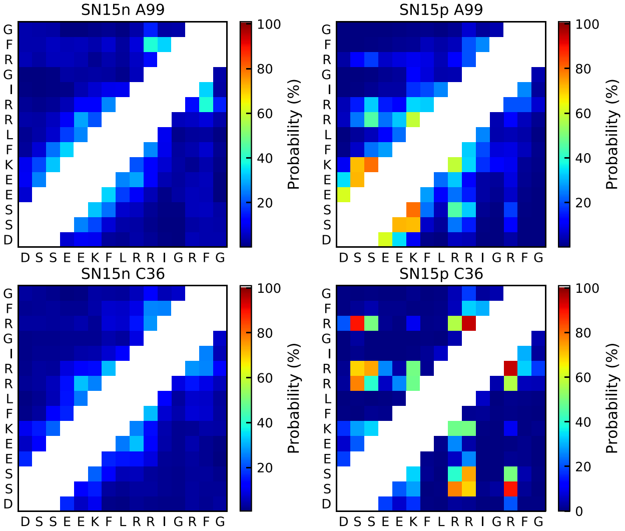


Figure 6. Contact map showing the probability of atoms in different residues being closer than 4 Å in SN15n (left column) and SN15p (right column) simulated with AMBER ff99SB-ILDN (top row) and CHARMM36m (bottom row). The two closest residues on each side as well as the residue itself are excluded from the analysis and therefore shown in white.

arginine and the phosphate group. Around 75% of the conformations also have the Arg9 close to pSer2, and 65% have Arg10 close. Approximately 90% of the conformations have Arg10 and Arg13 in close vicinity, which probably is because of both of them coordinating to a phosphate group simultaneously. The contact map also suggests that in many conformations all three arginines are coordinated simultaneously to the two phosphorylated serines. Such a coordination explains the presence of a bend in the middle of the peptide, as was shown by the DSSP analysis, and agrees with a more compact conformation with a smaller end-to-end distance. The high occurrence of such conformations can be explained by the strength of the interaction between the phosphate group and the guanidinium group of the arginine side chain, as this interaction has been shown to be of covalent-like stability.[20] However, this interaction appears stronger in the C36 force field than in A99, as the A99 simulation shows a lower amount of coordination between the arginines and the phosphate groups. The most probable close contact in the A99 simulation is between pSer3 and Lys6, and secondary between pSer2 and Glu5, as well as between pSer2 and Lys6. Overall, between residues 2 and 7, there is an enhanced contact between residues three or four steps away, compared to the C36 simulation. This is connected to the higher probability of helical structure, especially 3_{10}-helix, in this region of the peptide. Errington and Doig have shown that phosphorylation of a serine four neighbors away from a lysine stabilizes the α-helical structure through a strong interaction between the phosphate group and the positively charged lysine side chain. Having the lysine further away, the phosphorylation instead destabilizes the α-helix.[19] The enhanced contact between Ser2 and Lys6 upon phosphorylation is observed in both force fields, although the contact between Ser3 and Lys6 is preferential.

An analysis of the number of intrapeptide hydrogen bonds in each conformation (Figure 7) confirms that A99 shows more helical hydrogen bonds than C36. The C36 simulation depicts instead more hydrogen bonds between residues more than five neighbors away from each other, which is the category that the pSer−Arg hydrogen bonds fall into.

A comparison between the energy landscapes of the two force fields confirms the large differences between them, see Figure 9. First of all, C36 does exhibit a much smaller conformational landscape, in agreement with the narrower R_{ee} and R_g distributions. Although the A99 simulation has an overall higher content of helical structure in the N-terminal end, evident from Figures 4 and 5, this is not observed in the conformations of the energy minima. However, upon closer inspection of the other conformations in basin b_0 that are located very close to the minimum, it is revealed that some of them have a helical structure in the N-terminal part of the peptide, around the twist in the snapshot shown in Figure 9b of basin b_0, which corresponds to residues 4−6. Hence, it appears that the conformations with helical structure mostly fall within this basin. The most striking difference between the snapshots shown in Figure 9b,d is that almost all of the C36 conformations exhibit the same bend in the middle, whereas the two termini are allowed to point in more opposing directions for A99. The reason behind the bent structure in the C36 simulation is the electrostatic interaction and hydrogen bonding between the phosphate groups and arginine side chains as discussed above, which is shown in Figure 10. This figure also confirms that the same phosphate group can form hydrogen bonds with several arginines simultaneously, which was suggested by the contact map.

3.3. Effect of Phosphorylation and Experimental Comparison.

The end-to-end distance and the radius of gyration distributions (Figure 2) show that in both force fields, phosphorylation gives rise to a compaction of the peptide. On average, the radius of gyration is reduced from 9.99 ± 0.13 to 8.98 ± 0.12 Å for A99, and from 9.87 ± 0.08 to 8.13 ± 0.09 Å for

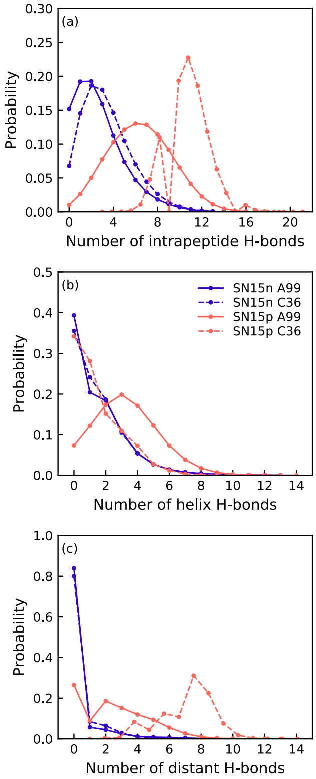


Figure 7. Probability distribution of the number of intrapeptide hydrogen bonds in total (a), associated with helices, that is, between residue *n* and *n* + *i*, where *i* = 3, 4, 5 (b), and between residue *n* and *n* + *i*, where *i* ≥ 6 (c). The legend applies to all panels.

C36. The increased compactness is also captured by the calculated SAXS spectra, shown in Figure 3 as a dimensionless Kratky plot.

To assess the performance of the force fields, experimental SAXS data were collected for the two peptides. The full concentration series are available in the Supporting Information, Figure S29. For SN15n, the different concentration curves agree, suggesting monomeric protein and no protein−protein interactions affecting the curves. The molecular weight, M_w, was calculated to be between 2.03 and 2.09 kDa for all the protein concentrations, see Table S1 in the Supporting Information, which is approximately 15% larger than the theoretical M_w of 1.80 kDa. Normally, the uncertainty of the molecular weight determined from SAXS is around 10%, but considering a higher uncertainty of the determined concentration in these measurements, 15% is acceptable. Hence, the SN15n peptide is regarded as monomeric and the collected SAXS data correspond to the form factor. In the case of the phosphorylated peptide, SN15p, the SAXS curves differ slightly with concentration; especially the forward scattering increases with concentration. This corresponds to an increase in molecular weight from 2.06 kDa at 1 mg/mL, to 3.01 kDa at 6 mg/mL (see Table S2), which compared to the theoretical value 1.96 kDa is 5 and 54% larger, respectively. Hence, this system shows some form of self-association with increased concentration, which is not unreasonable considering that phosphoserine can form strong interactions with arginines. However, since the molecular weight determined at 1 mg/mL is in good agreement with the theoretical value, these data are expected to correspond to the form factor. The form factor of SN15n and SN15p is presented in Figure 11a together with the form factors calculated from the simulations. It is shown that for the nonphosphorylated peptide both force fields agree with the experimental SAXS data, whereas for the phosphorylated peptide C36 is in disagreement, and possibly A99 as well. The radius of gyration of SN15n was determined to be 10.5 ± 0.2 or 9.9 ± 0.1 Å, using the pair distance distribution function or the Guinier approximation, respectively. Corresponding values for SN15p are 10.5 ± 0.2 and 9.6 ± 0.6 Å, respectively. Hence, the decrease in R_g upon phosphorylation observed in simulations is not supported by the experiments. However, from the upturn at low *q* for SN15p in Figure 11a, it is clear that the SAXS data show some aggregation of the sample, which might affect the full curve. Therefore, these data are less reliable than the SN15n data, and the R_g might be slightly overestimated. Nonetheless, it is unlikely that the effect would be large enough to make the simulations and experiments agree.

The Kratky plot in Figure 11b shows that there is good agreement in shape between the experiments and simulations for SN15n. For SN15p, the data are too noisy to state with certainty how the force fields compare. However, by comparing the simulations to the regularized curve fitted to the experimental data in the *P*(*r*) determination, a difference is observed as shown in Figure 12. The phosphorylated peptide is slightly less stiff/extended than the nonphosphorylated one, although the effect is much smaller than what both force fields predict. In addition, it appears that the self-association in the phosphorylated system has a minor effect on the shape. Hence, it appears reasonable to conclude that both force fields accurately capture the shape of the nonphosphorylated peptide, whereas they overestimate the compactness induced by phosphorylation. This is especially true for the C36 force field, but whereas A99 performs better, it is still not in agreement with the experimental data. Hence, neither of the force fields appears to accurately represent the phosphorylated residues.

Regarding the secondary structure, the two force fields give different responses to phosphorylation. For A99, there is an increase in the helical content in the N-terminal region, whereas C36 mostly shows an increase of bends. Experimental studies have shown that phosphoserine in the N-terminal position of a α-helix, or in an *i*, *i* + 4 position with lysine, stabilizes the α-helix,[18,19] which is the situation in the SN15p peptide. In the Ramachandran plot (Figure 5), a shift toward higher helical content for A99 is visible, whereas phosphorylation gives no clear effect in the Ramachandran plot for C36. Vymětal, Jurásková, and Vondrášek[39] recently published a study comparing the microscopic details, including conformational preferences, for terminally capped phosphorylated residues and their normal variants, obtained in three different force fields, including AMBER ff99SB with the phosaa10 parameters and CHARMM36m. They noticed that upon phosphorylation, the amount of extended and PPII-like conformations for serine decreased, whereas the amount of helical conformations increased, for the AMBER force field. The CHARMM force field on the other hand, showed decreasing amount of extended conformations, increasing amount of PPII-like conformations,

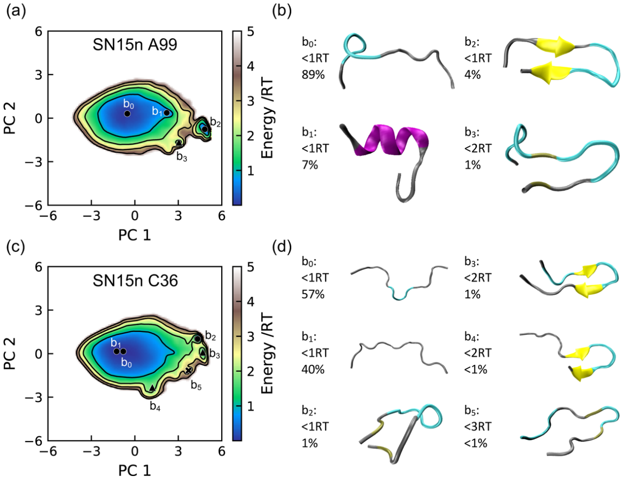


Figure 8. Energy landscape for SN15n using the first two principal components, and the conformation in each minimum from simulations with the AMBER ff99SB-ILDN (a,b) and CHARMM36m (c,d) force field. The energy landscapes were constructed using the same basis set, which makes them directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$. The minimum of each basin is represented by a marker: ●: energy $\leq 1RT$, ▲: $\leq 2RT$, and ×: $\leq 3RT$. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil), cyan is turn, magenta is α-helix, yellow is β-sheet, and tan is β-bridge. The N-terminus of each conformation is the leftmost/topmost end.

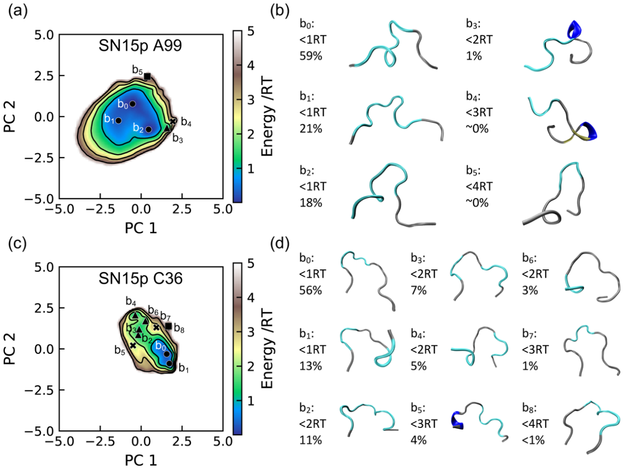


Figure 9. Energy landscape for SN15p using the first two principal components, and the conformation in each minimum from simulations with the AMBER ff99SB-ILDN (a,b) and CHARMM36m (c,d) force field. The energy landscapes were constructed using the same basis set, which makes them directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$. The minimum of each basin is represented by a marker: ●: energy $\leq 1RT$, ▲: $\leq 2RT$, ×: $\leq 3RT$, and ■: $\leq 4RT$. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil), cyan is turn, blue is 3_{10}-helix, and tan is β-bridge. The N-terminus of each conformation is the leftmost end.

and no significant changes in the helical content. Using the same classifications of extended, PPII-like and helical conformations for the serines in this study, the A99 force field gives a 20 percentage point decrease of extended conformations, a 32 percentage point decrease of PPII-like structures, and a 43 percentage point increase of helical conformations upon phosphorylation, in qualitative agreement with their study using the predecessor of the A99 force field. The C36 force field shows a 19 percentage point decrease of extended conformations, an 18 percentage point increase of PPII-like conformations, and a 15 percentage point increase of helical conformations upon phosphorylation, hence showing the same trend in extended and PPII-like structure as the study by Vymětal et al. Therefore, it appears that the force fields are

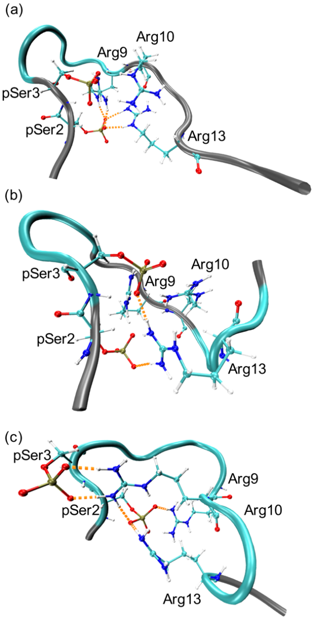


Figure 10. Snapshots of the peptide conformations in the minimum of basin b_0 (a), b_1 (b), and b_2 (c) in the energy landscape in Figure 9c. All atoms are shown in residues pSer2, pSer3, Arg9, Arg10, and Arg13. The orange dashed lines represent the hydrogen bonds between atoms in the side groups of the named residues.

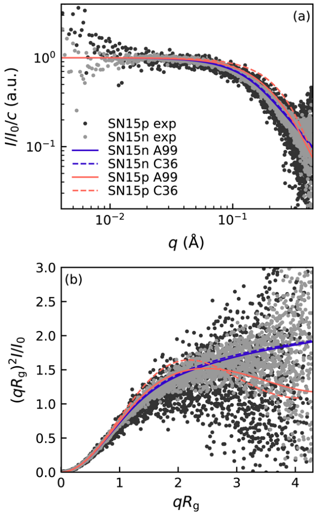


Figure 11. Form factor for SN15p and SN15n obtained by SAXS at 1.2 and ~4 mg/mL, respectively, at 20 °C, 150 mM NaCl, 20 mM Tris, and pH 7.5 shown as the scattering intensity (a) and a dimensionless Kratky plot (b). The simulated curves are included for comparison. The legend applies to both panels.

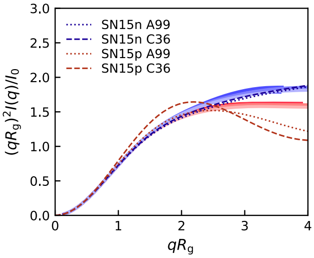


Figure 12. Dimensionless Kratky plot of the regularized curves fitted to the experimental SAXS data in the $P(r)$ determination, for the data obtained at 20 °C, 150 mM NaCl, 20 mM Tris, and pH 7.5. The blue solid lines correspond to SN15n, whereas the red solid lines correspond to SN15p. The color gradient shows the different concentrations, where the darkest color is the lowest concentration. The simulated curves are included for comparison.

biased toward different conformations of the phosphorylated residues, although the surrounding residues also affect the outcome.

For experimental reference, CD measurements were performed and the resulting spectra are shown in Figure 13.

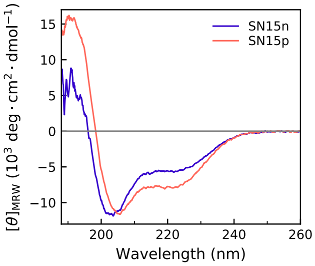


Figure 13. CD spectrum of SN15p and SN15n, measured at 20 °C in a 20 mM phosphate buffer at pH 7.5, with 150 mM NaF.

There is a clear difference between the nonphosphorylated and phosphorylated peptides, where phosphorylation induces a small shift of the global minimum towards higher wavelengths, a deeper secondary minimum around 222 nm, as well as a higher peak at 191 nm. All these changes are associated with an increase in α-helical structure. The experimental data are in qualitative agreement with measurements earlier performed by Raj et al.[11]

To achieve an assessment of the partition of secondary structural elements in the peptides, the data were analyzed with BeStSel. It is important to keep in mind that it is challenging to obtain highly accurate partitions from a CD spectrum, which is evident from different algorithms often giving different results. However, based on the quality of the fits to the experimental data, shown in Supporting Information Figure S30, and the normalized root mean square deviation being <0.02 for both peptides, the BeStSel results appears to be of adequate quality. The resulting secondary structure content is summarized in Table 2 and shows that the helical content is indeed increased upon phosphorylation, mostly at the expense of "others", which includes what in DSSP is classified as 3_{10}-helix, π-helix, bends, β-bridge, and irregular/loop. Also worth noticing is that the analysis suggests that the peptide contains a substantial amount

Table 2. Secondary Structure Content in the Phosphorylated and Nonphosphorylated Peptide at 150 mM NaF, pH 7.5, and 20 °C, According to the Analysis in BeStSel

	SN15n	SN15p
helix (%)	12.5	17.6
β-strand[a] (%)	18.6	23.9
turn (%)	16.8	15.4
others[b] (%)	52.3	43.2

[a] Antiparallel β-strand. [b] 3_{10}-helix, π-helix, bends, β-bridge, and irregular/loop.

(∼20%) of antiparallel β-strands, and that it increases upon phosphorylation.

Comparing the BeStSel results with the DSSP analysis of the simulations (Figure 4) is not straight-forward, because of the challenges of obtaining good estimates from the experimental data, as well as the experiments only providing an overall average and not information on residue level. Despite this, it is clear that the experimental analysis shows a higher structure content, since the simulations are highly dominated by irregular and bends, which is what BeStSel classifies as others. Hence, the simulations are not quantitatively comparable to the experiments regarding the secondary structure. However, the A99 force field captures an increase of helical structure upon phosphorylation. For β-strands, the content is low in all simulations, yet, as shown in Figures 8 and 9, β-strands and β-sheets do occur in local energy minima in all simulations except for the C36 simulation of SN15p. As these minima in some cases are separated by relatively high energy barriers, the sampling of them might however not be sufficient. This can be investigated further by employing enhanced sampling techniques. Overall, neither of the force fields shows an increase of β-sheetstructures on average upon phosphorylation.

For both force fields, the presence of phosphorylated residues gives rise to specific interactions between certain amino acids, shown in the contact map in Figure 6. It is clear that it is the phosphorylated residues that are involved in the more prominent contacts and mainly with positively charged residues. The effect is especially large for C36, where phosphorylation causes highly conserved contacts between pSer and Arg. These contacts are formed by hydrogen bonds in addition to electrostatic interaction, which explains why there is an increase of intrapeptide hydrogen bonds after phosphorylation (Figure 7). This increase is observed for both force fields, whereas Figure 7b,c reveals that the type of hydrogen bonds differs. The C36 force field gives the same number of hydrogen bonds associated with helices (i.e., bonds between a residue and another residue three to five residues away) for the phosphorylated as well as the nonphosphorylated peptides, whereas the number of hydrogen bonds between more distant residues is much higher in the phosphorylated case. Hydrogen bonds between pSer and Arg all fall into this last category. The A99 force field shows an increase in both the number of helix hydrogen bonds and hydrogen bonds between more distant residues upon phosphorylation. The increase in helix hydrogen bonds is related to the increase in helical content.

Overall, the agreement between the experimental data and the simulations are worse for the C36 force field than the A99 force field. The main difference between the force fields is attributed to the highly conserved contacts between the phosphorylated serines and arginines in the C36 simulation, causing more compact and less interchangeable conformations. Hence, it appears that the interaction between phosphate and arginine, even though experiments have shown that it can be exceptionally strong for an intermolecular interaction,[20] is too strong in this force field and that A99 gives a better representation of this peptide. However, the agreement is still not satisfactory, showing the need for new parameterizations of phosphorylated amino acids in force fields suitable for IDPs.

4. CONCLUSIONS

AMBER ff99SB-ILDN with the TIP4P-D water model and CHARMM36m with the CHARMM-modified TIP3P water model give overall similar results for the SN15n peptide. Differences were only observed in the secondary structure, where A99 gave a larger content of 3_{10}-helix than C36. Both force fields showed great agreement with the experimental SAXS data, whereas experimental CD data suggested a higher structure content than what was observed in the simulations. Therefore, it is concluded that both force fields are in experimental agreement regarding size and shape, whereas improvements can be made regarding capturing the secondary structure.

In the simulation of the phosphorylated peptide, SN15p, the AMBER force field was complemented with phosaa10 parameters for the phosphorylated serines, whereas this was already included in the CHARMM force field. Both force fields showed a compaction of the peptide compared to the nonphosphorylated peptide, but this effect was further enhanced for C36, in which multiple hydrogen bonds between the phosphate groups and arginines trapped the peptide in more bent conformations. A99 gave an increase of helical content in the N-terminal part of the peptide upon phosphorylation, whereas C36 showed no differences in structure content. CD data showed an overall increase in both α-helical and β-strand content, therefore suggesting that A99 qualitatively can capture some of the aspects of phosphorylation in IDPs, while still giving too compact conformations. Hence, revision of the parameters for phosphorylated residues is encouraged for both force fields.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b01190.

Detailed assessment of the simulation convergence and sampling, results on salt effects in the simulations, additional SAXS data, and the BeStSel fit to the experimental CD data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ellen Rieloff − *Division of Theoretical Chemistry, Lund University, SE-221 00 Lund, Sweden;* orcid.org/0000-0002-4502-8395; Email: ellen.rieloff@teokem.lu.se

Marie Skepö − *Division of Theoretical Chemistry, Lund University, SE-221 00 Lund, Sweden; LINXS—Lund Institute of Advanced Neutron and X-ray Science, SE-223 70 Lund, Sweden;* orcid.org/0000-0002-8639-9993; Email: marie.skepo@teokem.lu.se

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.9b01190

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Samuel Lenton for performing the SAXS experiments at beamline B21, Diamond Light Source, UK. Financial support has been given by the Royal Physiographic Society of Lund, the Magnus Bergvall's foundation, and the Bertil Andersson's foundation. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Center for Scientific and Technical Computing at Lund University (Lunarc).

REFERENCES

(1) Keith Dunker, A.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z. Intrinsically disordered protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26−59.

(2) Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527−533.

(3) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635−645.

(4) Liu, J.; Faeder, J. R.; Camacho, C. J. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19819−19823.

(5) Keith Dunker, A.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **2002**, *41*, 6573−6582.

(6) Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Keith Dunker, A. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004**, *32*, 1037−1049.

(7) Gao, J.; Xu, D. Correlation Between Posttranslational Modification and Intrinsic Disorder in Protein. In *Biocomputing 2012*; Altman, R. B., Keith Dunker, A., Hunter, L., Murray, T. A., Klein, T. E., Eds.; World Scientific Publishing Co. Pte. Ltd., 2012; pp 94−103.

(8) Boskey, A. L. Phosphoproteins and Biomineralization. *Phosphorus, Sulfur Silicon Relat. Elem.* **1999**, *144*, 189−192.

(9) Moreno, E. C.; Zahradnik, R. T. Demineralization and Remineralization of Dental Enamel. *J. Dent. Res.* **1979**, *58*, 896−903.

(10) Hay, D. I.; Smith, D. J.; Schluckebier, S. K.; Moreno, E. C. Basic Biological Sciences Relationship between Concentration of Human Salivary Statherin and Inhibition of Calcium Phosphate Precipitation in Stimulated Human Parotid Saliva. *J. Dent. Res.* **1984**, *63*, 857−863.

(11) Raj, P. A.; Johnsson, M.; Levine, M. J.; Nancollas, G. H. Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization. *J. Biol. Chem.* **1992**, *267*, 5968−5976.

(12) Little, E. M.; Holt, C. An equilibrium thermodynamic model of the sequestration of calcium phosphate by casein phosphopeptides. *Eur. Biophys. J.* **2004**, *33*, 435−447.

(13) Holt, C.; Lenton, S.; Nylander, T.; Sørensen, E. S.; Teixeira, S. C. M. Mineralisation of soft and hard tissues and the stability of biofluids. *J. Struct. Biol.* **2014**, *185*, 383−396.

(14) Lenton, S.; Grimaldo, M.; Roosen-Runge, F.; Schreiber, F.; Nylander, T.; Clegg, R.; Holt, C.; Härtlein, M.; García Sakai, V.; Seydel, T.; Marujo Teixeira, S. C. Effect of Phosphorylation on a Human-like Osteopontin Peptide. *Biophys. J.* **2017**, *112*, 1586−1596.

(15) Gong, C.-X.; Iqbal, K. Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease. *Curr. Med. Chem.* **2008**, *15*, 2321−2328.

(16) Johnson, L. N.; Lewis, R. J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **2001**, *101*, 2209−2242.

(17) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **2011**, *1*, 90.

(18) Andrew, C. D.; Warwicker, J.; Jones, G. R.; Doig, A. J. Effect of Phosphorylation on α-Helix Stability as a Function of Position. *Biochemistry* **2002**, *41*, 1897−1905.

(19) Errington, N.; Doig, A. J. A Phosphoserine−Lysine Salt Bridge within an α-Helical Peptide, the Strongest α-Helix Side-Chain Interaction Measured to Date. *Biochemistry* **2005**, *44*, 7553−7558.

(20) Woods, A. S.; Ferré, S. Amazing Stability of the Arginine−Phosphate Electrostatic Interaction. *J. Proteome Res.* **2005**, *4*, 1397−1402.

(21) Mandell, D. J.; Chorny, I.; Groban, E. S.; Wong, S. E.; Levine, E.; Rapp, C. S.; Jacobson, M. P. Strengths of Hydrogen Bonds Involving Phosphorylated Amino Acid Side Chains. *J. Am. Chem. Soc.* **2007**, *129*, 820−827.

(22) Rauscher, S.; Pomès, R. Molecular simulations of protein disorder. *Biochem. Cell Biol.* **2010**, *88*, 269−290.

(23) Burger, V.; Gurry, T.; Stultz, C. Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **2014**, *6*, 2684−2719.

(24) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10*, 5113−5124.

(25) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.

(26) Henriques, J.; Cragnell, C.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 3420−3431.

(27) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513−5524.

(28) Henriques, J.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the TIP4P-D Water Model and the Representativeness of Protein Disorder Models. *J. Chem. Theory Comput.* **2016**, *12*, 3407−3415.

(29) Naqvi, M. A.; Rauscher, S.; Pomès, R.; Rousseau, D. The Conformational Ensemble of the β-Casein Phosphopeptide Reveals Two Independent Intrinsically Disordered Segments. *Biochemistry* **2014**, *53*, 6402−6408.

(30) Stanley, N.; Esteban-Martín, S.; De Fabritiis, G. Kinetic modulation of a disordered protein domain by phosphorylation. *Nat. Commun.* **2014**, *5*, 5272.

(31) Karim, C. B.; Michel Espinoza-Fonseca, L.; James, Z. M.; Hanse, E. A.; Gaynes, J. S.; Thomas, D. D.; Kelekar, A. Structural Mechanism for Regulation of Bcl-2 protein Noxa by phosphorylation. *Sci. Rep.* **2015**, *5*, 14557.

(32) Ithuralde, R. E.; Turjanski, A. G. Phosphorylation Regulates the Bound Structure of an Intrinsically Disordered Protein: The p53-TAZ2 Case. *PLoS One* **2016**, *11*, No. e0144284.

(33) Ilizaliturri-Flores, I.; Correa-Basurto, J.; Bello, M.; Rosas-Trigueros, J. L.; Zamora-López, B.; Benítez-Cardoza, C. G.; Zamorano-Carrillo, A. Mapping the intrinsically disordered properties of the flexible loop domain of Bcl-2: a molecular dynamics simulation study. *J. Mol. Model.* **2016**, *22*, 98.

(34) Colson, B. A.; Thompson, A. R.; Espinoza-Fonseca, L. M.; Thomas, D. D. Site-directed spectroscopy of cardiac myosin-binding protein C reveals effects of phosphorylation on protein structural dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 3233−3238.

(35) Gandhi, N. S.; Kukic, P.; Lippens, G.; Mancera, R. L. Molecular Dynamics Simulation of Tau Peptides for the Investigation of Conformational Changes Induced by Specific Phosphorylation Patterns. In *Tau Protein: Methods and Protocols*; Smet-Nocca, C., Ed.; Springer New York: New York, NY, 2017; pp 33−59.

(36) Hendus-Altenburger, R.; Lambrughi, M.; Terkelsen, T.; Pedersen, S. F.; Papaleo, E.; Lindorff-Larsen, K.; Kragelund, B. B. A phosphorylation-motif for tuneable helix stabilisation in intrinsically

disordered proteins - Lessons from the sodium proton exchanger 1 (NHE1). *Cell. Signalling* **2017**, *37*, 40−51.

(37) Luo, M.; Gao, Y.; Yang, S.; Quan, X.; Sun, D.; Liang, K.; Li, J.; Zhou, J. Computer simulations of the adsorption of an N-terminal peptide of statherin, SN15, and its mutants on hydroxyapatite surfaces. *Phys. Chem. Chem. Phys.* **2019**, *21*, 9342−9351.

(38) Mao, C. M.; Sampath, J.; Sprenger, K. G.; Drobny, G.; Pfaendtner, J. Molecular Driving Forces in Peptide Adsorption to Metal Oxide Surfaces. *Langmuir* **2019**, *35*, 5911−5920.

(39) Vymětal, J.; Juráskova, V.; Vondrášek, J. AMBER and CHARMM Force Fields Inconsistently Portray the Microscopic Details of Phosphorylation. *J. Chem. Theory Comput.* **2019**, *15*, 665−679.

(40) 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* **2006**, *65*, 712−725.

(41) Homeyer, N.; Horn, A. H. C.; Lanig, H.; Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281−289.

(42) Steinbrecher, T.; Latzer, J.; Case, D. A. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory Comput.* **2012**, *8*, 4405−4412.

(43) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(44) Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. Forcefield_PTM: Ab Initio Charge and AMBER Forcefield Parameters for Frequently Occurring Post-Translational Modifications. *J. Chem. Theory Comput.* **2013**, *9*, 5653−5674.

(45) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2016**, *14*, 71−73.

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

(47) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(48) Schrödinger, LLC. *The PyMOL Molecular Graphics System*, version 1.2r1, 2009.

(49) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminf.* **2012**, *4*, 17.

(50) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43−56.

(51) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(52) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(53) Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. *Solving Software Challenges for Exascale*; Springer: Cham, 2015; pp 3−27.

(54) 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.

(55) Berendsen, H.; Van Gunsteren, W. *Molecular-Dynamics Simulations of Statistical-Mechanical Systems*; North-Holland, 1986; pp 43−65.

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

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

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

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

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

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

(62) Svergun, D.; Barberato, C.; Koch, M. H. J. CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768−773.

(63) Campos, S. R. R.; Baptista, A. M. Conformational Analysis in a Multidimensional Energy Landscape: Study of an Arginylglutamate Repeat. *J. Phys. Chem. B* **2009**, *113*, 15989−16001.

(64) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(65) Stone, J. E. An Efficient Library for Parallel Ray Tracing and Animation. M.Sc. Thesis, Computer Science Department, University of Missouri-Rolla, 1998.

(66) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.

(67) Mihalyi, E. Numerical values of the absorbances of the aromatic amino acids in acid, neutral, and alkaline solutions. *J. Chem. Eng. Data* **1968**, *13*, 179−182.

(68) Franke, D.; Petoukhov, M. V.; Konarev, P. V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H. D. T.; Kikhney, A. G.; Hajizadeh, N. R.; Franklin, J. M.; Jeffries, C. M.; Svergun, D. I. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **2017**, *50*, 1212−1225.

(69) Kuipers, B. J. H.; Gruppen, H. Prediction of Molar Extinction Coefficients of Proteins and Peptides Using UV Absorption of the Constituent Amino Acids at 214 nm To Enable Quantitative Reverse Phase High-Performance Liquid Chromatography−Mass Spectrometry Analysis. *J. Agric. Food Chem.* **2007**, *55*, 5445−5451.

(70) Micsonai, A.; Wien, F.; Kernya, L.; Lee, Y.-H.; Goto, Y.; Réfrégiers, M.; Kardos, J. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E3095−E3103.

(71) Micsonai, A.; Wien, F.; Bulyáki, É.; Kun, J.; Moussong, É.; Lee, Y.-H.; Goto, Y.; Réfrégiers, M.; Kardos, J. BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res.* **2018**, *46*, W315−W322.

(72) Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J. Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **2019**, *10*, 2227−2234.

Supporting information for: Phosphorylation of a disordered peptide – structural effects and force field inconsistencies

Ellen Rieloff*,† and Marie Skepö*,†,‡

†*Division of Theoretical Chemistry, Lund University, POB 124, SE-221 00 Lund, Sweden*

‡*LINXS – Lund Institute of Advanced Neutron and X-ray Science, Scheelevägen 19, SE-223 70 Lund, Sweden*

E-mail: ellen.rieloff@teokem.lu.se; marie.skepo@teokem.lu.se

1 Convergence and sampling in the simulations

Assessing the sampling quality and uncertainty of the simulations is important for ensuring reliable result. In this study each system has been simulated in five replicates that have been combined to a single trajectory before performing the final analysis. In this section the five replicates are compared to assess sampling quality, by visual analysis of the energy landscapes obtained from principal component analysis (PCA), as well as by the time evolution and probability distribution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g). Furthermore, error estimates of R_{ee} and R_g from the concatenated simulation trajectory have been obtained by block averaging, and the sampling has been assessed by observing the auto-correlation function. Below the results are presented for each system.

1.1 SN15n with AMBER ff99SB-ILDN

The energy landscapes obtained from PCA analysis using the first two principal components, presented in Figure S1, are overall similar in shape for all replicates. Hence, apart from the third replicate, the replicates appear to be sampling the same conformational space. Regarding the third replicate, the lowest energy minimum is located in a narrow basin in the right side of the plot, which is an area not sampled in the others. Apart from this minimum the conformational landscape is similar in distribution to the other replicates.

From the time evolution of the end-to-end distance and the radius of gyration, presented in Figure S2, it is suggested that the low-energy basin in the third replicate contains more compact conformations than otherwise sampled. For the replicates overall, R_{ee} and R_g

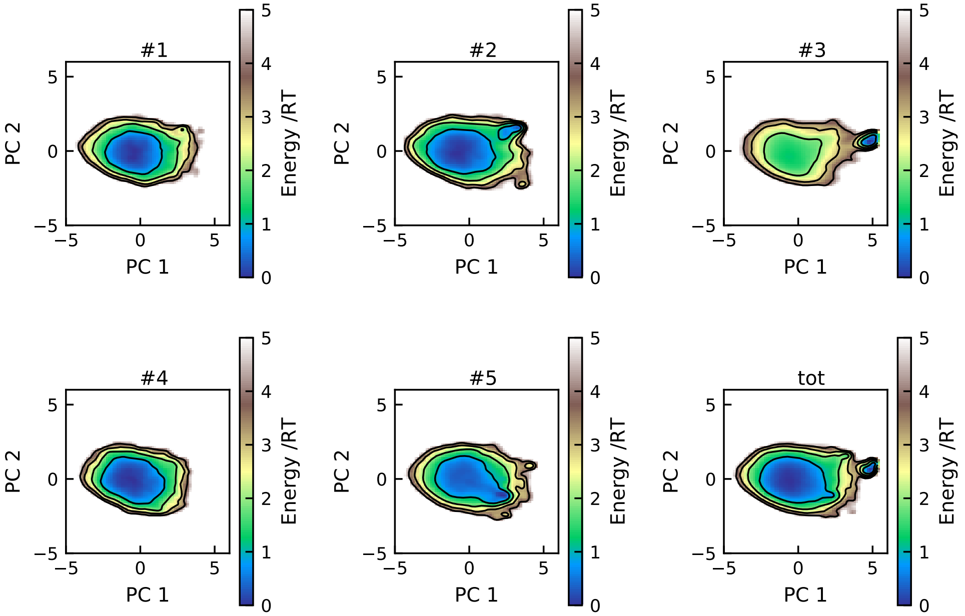


Figure S1: Energy landscapes for the five replicates and the concatenated trajectory of SN15n in AMBER ff99SB-ILDN, using the two first principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$.

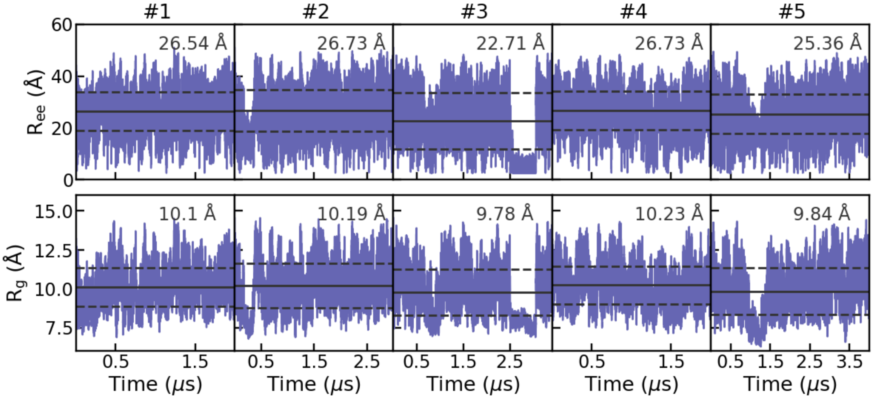


Figure S2: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of SN15n using AMBER ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

changes rapidly during the simulations, which is expected for intrinsically disordered proteins. However, some replicates, especially the third one, show time periods of a drastically lower R_g and R_{ee}, which suggests that the peptide can temporary get stuck in a more compact folded structure. The relatively long period of compact conformations in the third replicate results in a lower mean R_{ee} and R_g than the other replicates, although still within the standard deviation of the others. These compact conformations give rise to a peak at lower values in the density distributions of R_{ee} and R_g, presented in Figure S3. Apart from this peak, there is an overall good agreement between the density distributions of the different replicates. Replicate two and five also show some minor peaks at low values, which is expected from Figure S2.

The more compact conformations that occur in replicate number two, three, and five, are shown in Figure S4. It is clear that these conformations contain more secondary structure than the majority in the ensemble. Since the energy landscapes in Figure S1 contain secondary basins with low minimum energy for these replicates, it is suggested that these folded structures also have low energy, but are separated from the primary unstructured basin by

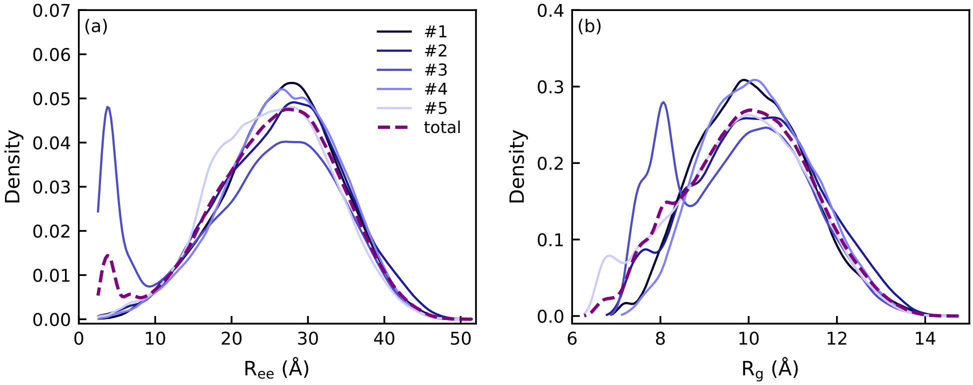


Figure S3: Density estimates of the end-to-end distance, R_{ee} (a), and the radius of gyration, R_g (b), for the five replicates in the simulation of SN15n using AMBER ff99SB-ILDN, obtained from a Gaussian kernel estimator. The dashed purple line shows the distribution for the five replicates combined.

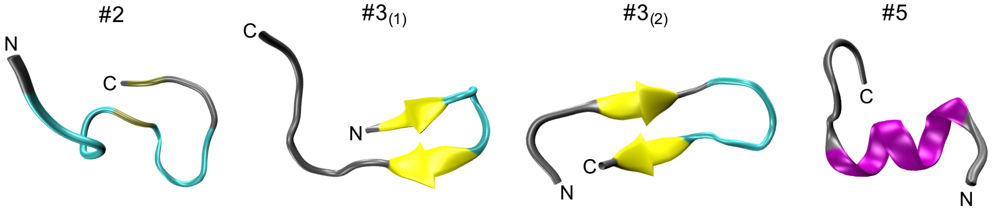


Figure S4: Snapshots representing the compact conformations identified from the end-to-end distance and radius of gyration time evolution of the simulations of SN15n using AMBER ff99SB-ILDN. The numbers above state the replicate where they occur. The peptide is colored according to secondary structure, as determined by VMD. Silver corresponds to irregular (coil), cyan to turn, tan to β-bridge, yellow to β-sheet, and magenta to α-helix. The N- and C-termini are marked with N or C, respectively.

energy barriers of different heights. Therefore, it still remains uncertain whether these more compact conformations are over-represented or not sampled enough. However, they appear to have relatively small impact on the overall ensemble, such that this simulation still can be compared qualitatively to the other force field and sequence.

Figure S5 shows the autocorrelation function and the standard error as a function of the block length, for the end-to-end distance and radius of gyration in the concatenated

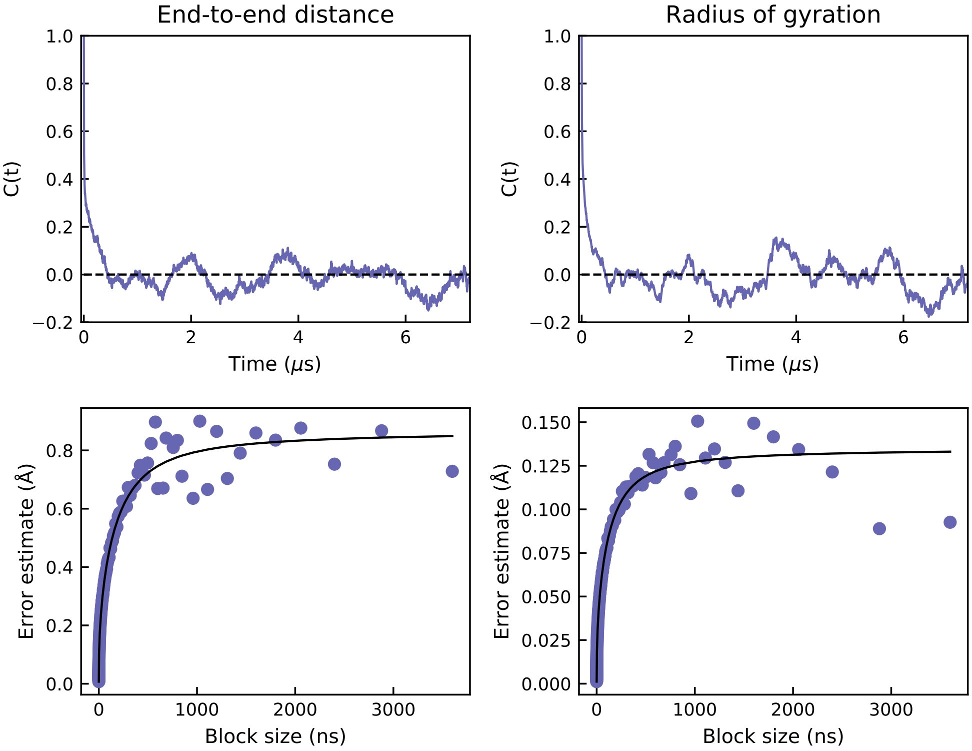


Figure S5: Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (left) and the radius of gyration (right) for the concatenated simulation of SN15n using AMBER ff99SB-ILDN.

trajectory of simulation length 14.4 µs. For both R_{ee} and R_g, the autocorrelation decreases to zero within approximately 0.5 µs, although with time it fluctuates around zero and some smaller peaks are visible. However, the error estimate converges with increased block size, suggesting that these observables have been sufficiently sampled. Hence, it appears that the sampling is not completely satisfying, although adequate for obtaining a representation of the dominating structures of the conformational ensemble. However, it is important to keep in mind that exact numbers can change upon further sampling, although the trends observed when comparing to the other simulations are expected to remain.

1.2 SN15n with CHARMM36m

The energy landscapes for the different replicates show the same overall distribution, see Figure S6. The first three replicates also display the same secondary basin in the right side of the plot. In analogy to the previously discussed force field, these are probably more folded conformations. The fourth replicate also displays a separated basin, but in another part of the plot. However, overall it appears that all replicates have sampled approximately the same conformational space.

The narrow basins observed in Figure S6 appear to once again be related to short periods of drastically lower R_{ee} and R_g, as shown in Figure S7. The third replicate has both a mean R_{ee} and R_g slightly smaller than the others, although still equal within the standard

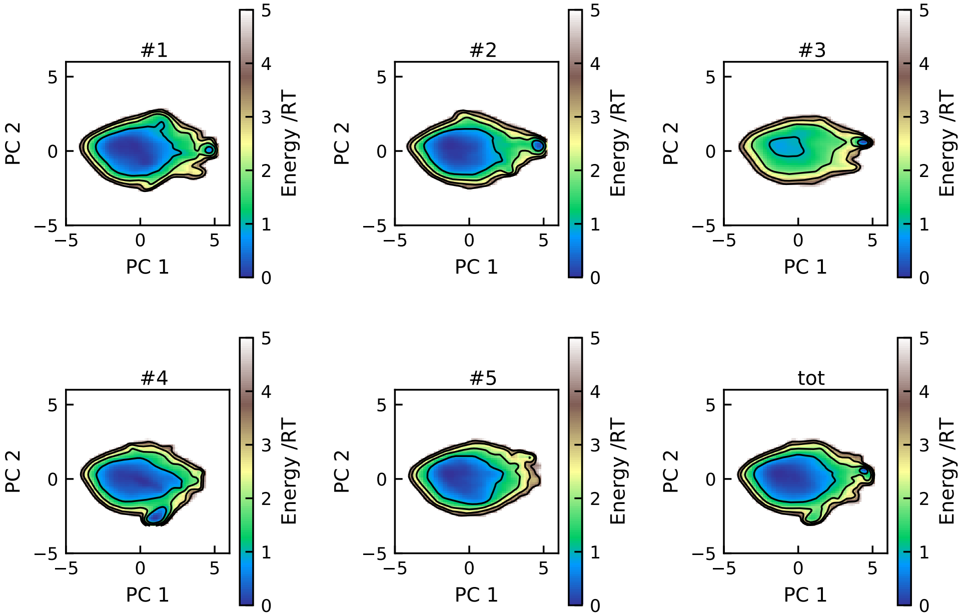


Figure S6: Energy landscapes for the five replicates and the concatenated trajectory of SN15n in CHARMM36m, using the two first principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$.

deviation. This is not surprising as it was this replicate that showed the largest deviation from the others in Figure S6.

Overall, for the probability distribution of the end-to-end distance shown in Figure S8a, there is good agreement between the replicates. However, the corresponding plot for the

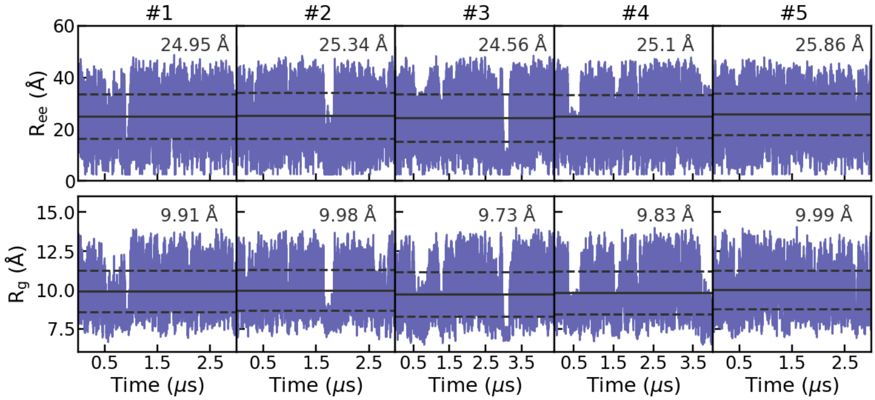


Figure S7: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of SN15n using CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

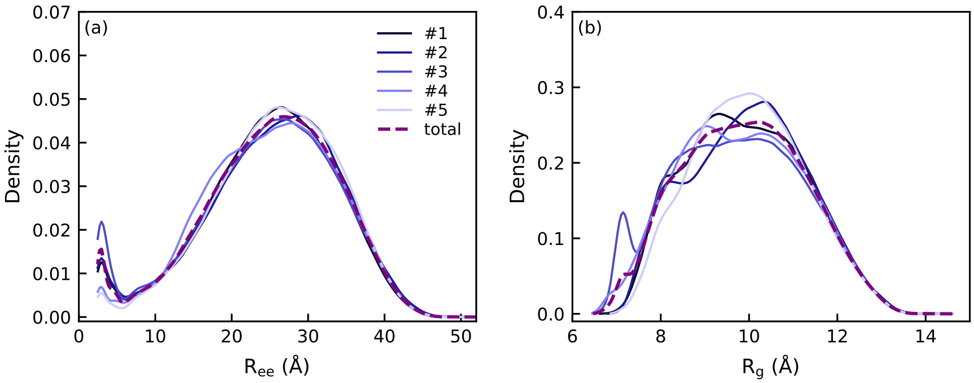


Figure S8: Density estimates of the end-to-end distance, R_{ee} (a), and the radius of gyration, R_g (b), for the five replicates in the simulation of SN15n using CHARMM36m, obtained from a Gaussian kernel estimator. The dashed purple line shows the distribution for the five replicates combined.

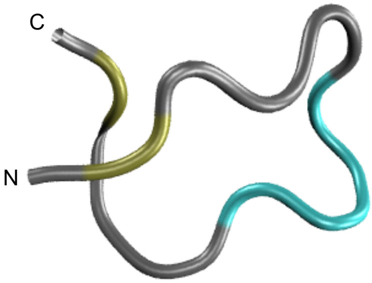


Figure S9: Representative snapshot of the conformation with low R_{ee} and R_g in the third replicate in the simulation of SN15n with CHARMM36m. The peptide is colored according to secondary structure, as determined by VMD. Silver corresponds to irregular (coil), cyan to turn, and tan to β-bridge. The N and C mark the corresponding terminus.

radius of gyration (Figure S8b) shows more variation. The main difference appears from the third replicate that contains a well-defined secondary peak around 7 Å. This peak corresponds to the conformations around 3-3.2 μs which display R_g values well below the average. A representative snapshot of this range shown in Figure S9 reveals that both ends of the peptide are close together, exhibiting β-strand formation near the end parts of the peptide. However, overall this conformation appears to have a limited influence on the total simulation, as shown by the rather smooth distributions of R_{ee} and R_g when combining all the replicates (Figure S8).

The correlation functions for both the R_{ee} and the R_g decrease to zero within 0.5 μs for the concatenated trajectory, and afterwards they fluctuates around zero with no major correlation, see Figure S10. This suggests that the total simulation of 17.4 μs is long enough. The error estimate from the block analysis, presented in the same figure, points towards the same conclusion.

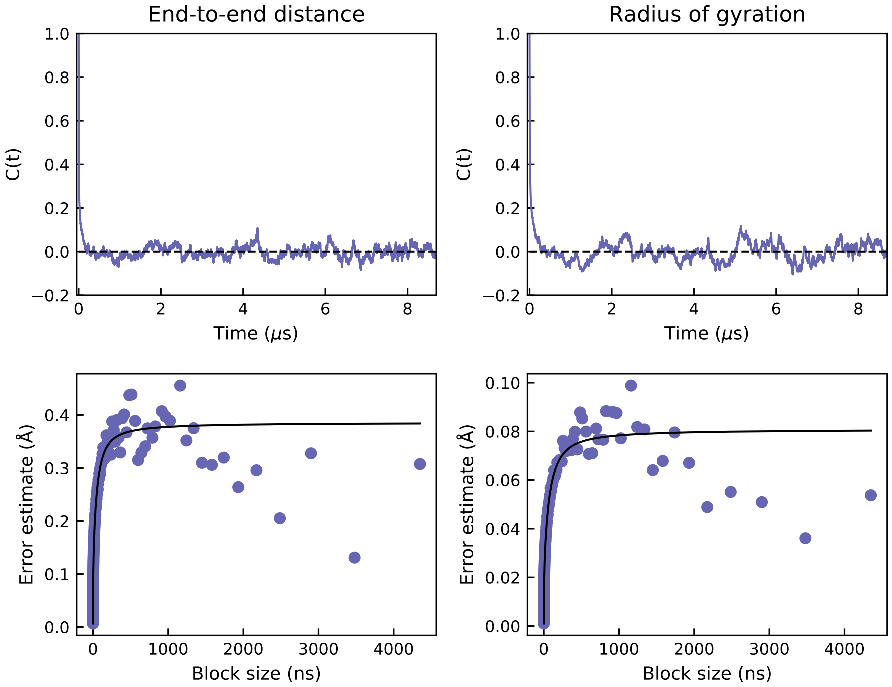


Figure S10: Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (left) and the radius of gyration (right) for the concatenated simulation of SN15n using CHARMM36m.

1.3 SN15p with AMBER ff99SB-ILDN

For the phosphorylated peptide, simulated in AMBER ff99SB-ILDN, the energy landscapes of the five replicates (Figure S11) once again show equal distribution, although the first and third replicate display smaller areas of energy $\leq 1RT$ than the others. From the time evolution of the R_{ee} and the R_g in Figure S12 it is clear that these two replicates posses smaller mean values than the others, although equal within the standard deviation. Both show clear periods of at least 1 μs in which the radius of gyration is distinctly lowered. These regions mainly consist of conformations with some β-strand structure, such as exemplified in Figure S13.

The density distribution of the end-to-end distance, presented in Figure S14, reveals two

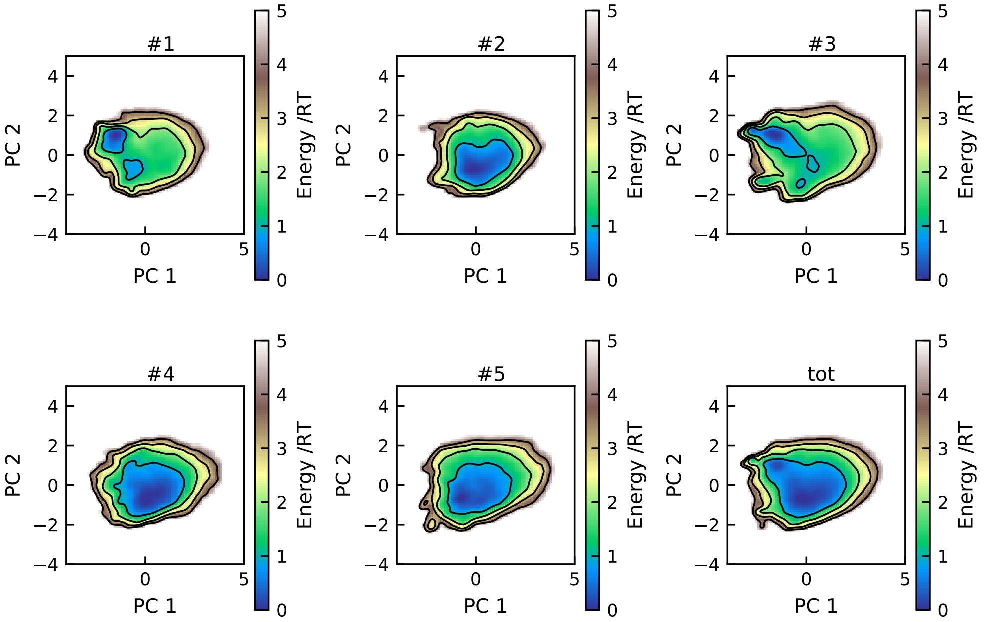


Figure S11: Energy landscapes for the five replicates and the concatenated trajectory of SN15p in AMBER ff99SB-ILDN, using the two first principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$.

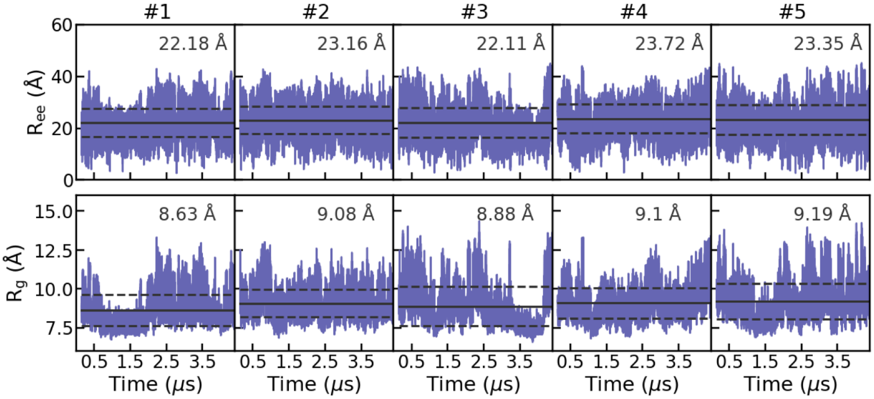


Figure S12: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of SN15p using AMBER ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

different groups: replicate one and three are in good agreement, while the remaining ones also show good agreement with each other. The total simulation, of all replicates combined, still shows a smooth distribution. The radius of gyration density distribution, displayed in the same figure, shows an even larger variety. The peak positions are once again split into

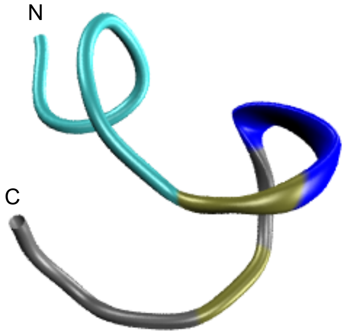


Figure S13: Representative snapshot of the more compact conformation in the first and third replicate in the simulation of SN15p with AMBER ff99SB-ILDN. The peptide is colored according to secondary structure, as determined by VMD. Cyan corresponds to turn, tan to β-bridge, blue to 3_{10}-helix, and silver to irregular (coil). The N and C mark the corresponding terminus.

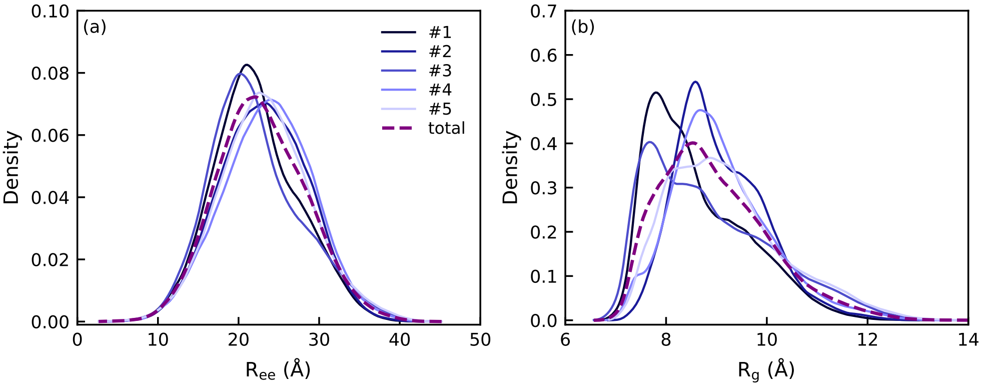


Figure S14: Density estimates of the end-to-end distance, R_{ee} (a), and the radius of gyration, R_g (b), for the five replicates in the simulation of SN15p using AMBER ff99SB-ILDN, obtained from a Gaussian kernel estimator. The dashed purple line shows the distribution for the five replicates combined.

the same two groups; where the first and third have their main peaks at a smaller radius of gyration. Still, the density distribution of the total simulation is relatively smooth, with an elongated tail at larger values.

Focusing on the concatenated trajectory of all the replicates together, the autocorrelation function of the end-to-end distance (Figure S15) decreases to zero within 0.5 µs, and stays close to zero afterwards. The autocorrelation of the radius of gyration first reaches zero after approximately 1 µs, but continues to oscillate in a regular pattern, suggesting that some correlation still remains. However, the error estimates from block analysis, shown in the same figure, converges nicely, suggesting a sufficiently long simulation (21.25 µs).

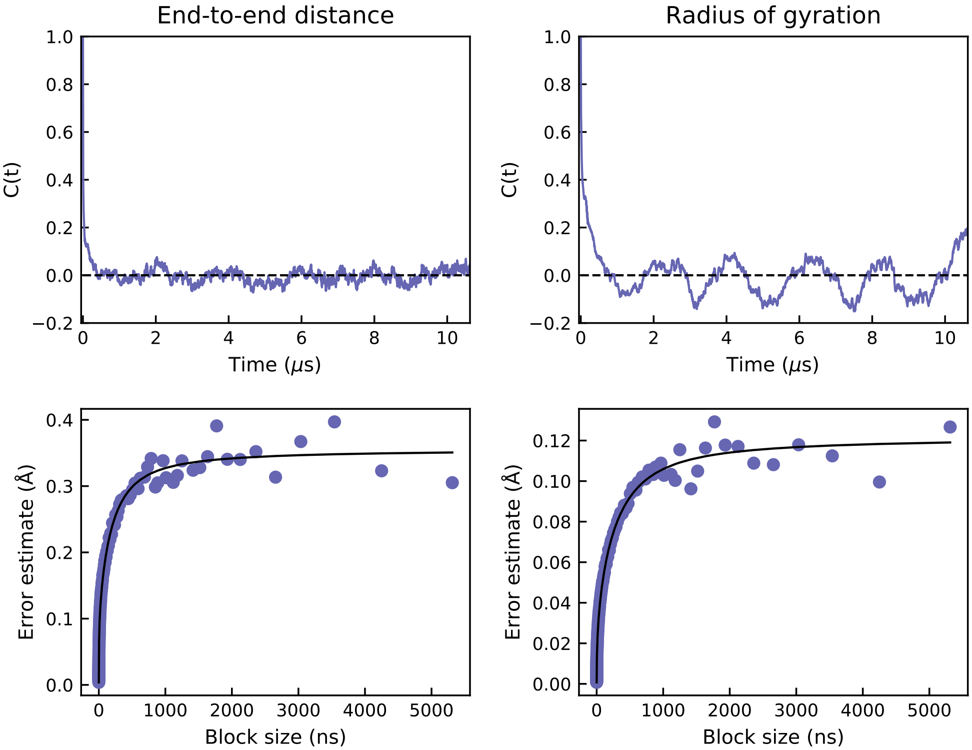


Figure S15: Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (left) and the radius of gyration (right) for the concatenated simulation of SN15p using AMBER ff99SB-ILDN.

1.4 SN15p with CHARMM36m

The SN15p peptide simulated with CHARMM36m shows the most diversity in the energy landscape between different replicates, as seen in Figure S16. This suggests that the simulation is not sampled well enough. In addition, compared to the other conditions, the area of the energy landscape is smaller, suggesting that there is less variation in the conformational ensemble. Indeed, the time evolutions of R_{ee} and R_g (Figure S17) show smaller spread compared to AMBER ff99SB-ILDN. Neither of the replicates show any periods of drastically different R_{ee}, except for replicate four, which exhibits a decrease in the end of the trajectory. The R_g on the other hand shows some different plateau values in the first, fourth, and fifth replicate. However, the difference is too small to give noticeable differences in the trajectory

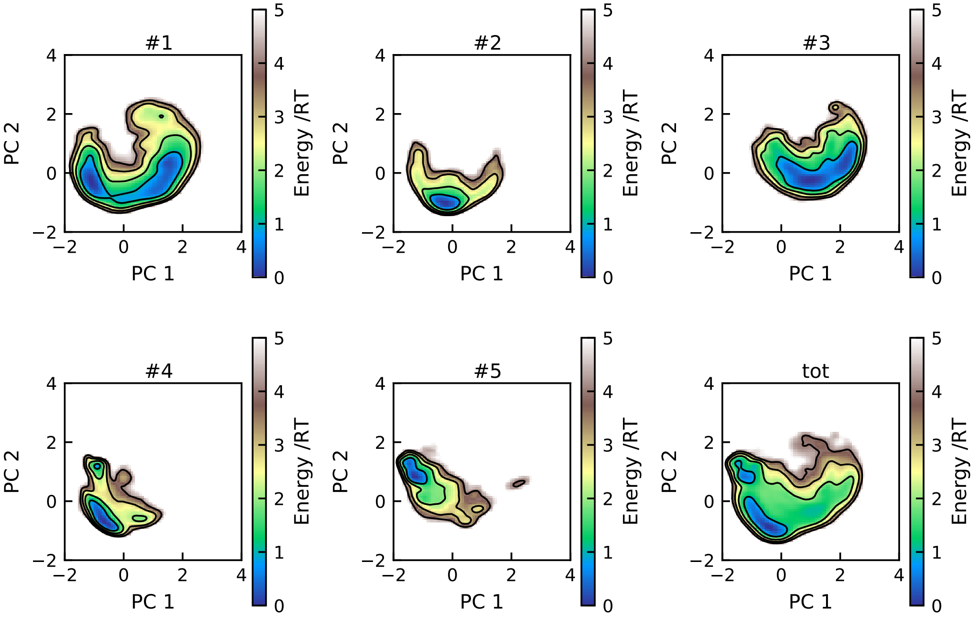


Figure S16: Energy landscapes for the five replicates and the concatenated trajectory of SN15p in CHARMM36m, using the two first principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$.

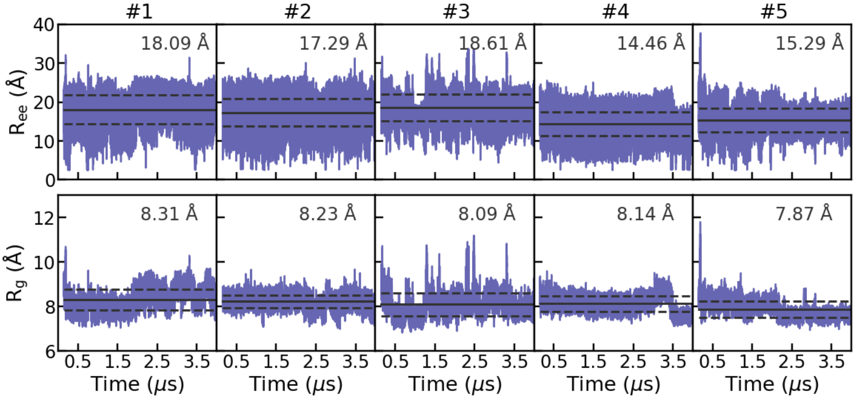


Figure S17: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of SN15p using CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

upon visual inspection. In contrary to the previously described simulations, there are no regions showing a significant amount of secondary structure.

The rather narrow density distributions of the end-to-end distance and radius of gyration, shown in Figure S18, confirm the smaller conformational ensemble, compared to the other force field. For the R_{ee}, the replicates are arranged in two groups; the first consisting of the first three replicates and the second one of the last two replicates. Within the groups there is good agreement, while the peak values of the two groups are separated by approximately 7 Å. The density distribution of the concatenated simulation shows a rather smooth distribution with peak value in between the two groups. For the radius of gyration, the five replicates display approximately the same range, however, different distributions within the range. Replicate four and five display a bimodal distribution, while the second replicate shows a smooth gaussian-like distribution coinciding with one of the peaks in the bimodal distribution. The remaining replicates show a “peak with shoulder”-distribution, where the first replicate has a significant shoulder at higher values than what is sampled in the other replicates. Overall, this suggests that more sampling might be required for obtaining reliable

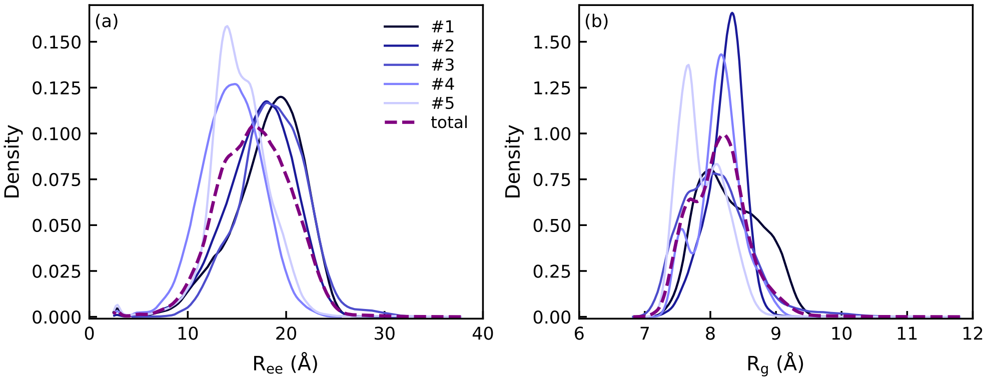


Figure S18: Density estimates of the end-to-end distance, R_{ee} (a) and the radius of gyration, R_g (b), for the five replicates in the simulation of SN15p using CHARMM36m, obtained from a Gaussian kernel estimator. The dashed purple line shows the distribution for the five replicates combined.

results, although it appears that the actual average values will not change drastically.

The autocorrelation function and error estimate from block averaging for R_{ee} and R_g of the total simulation is presented in Figure S19. It appears that the correlation time is rather long for the R_{ee}, as the autocorrelation function reaches zero first after almost 4 µs. In addition, the error estimate is not fully converged either, which together suggest the need of longer simulation time. The autocorrelation function of the R_g decreases faster to zero, although still relatively slow. However, the error estimate appears to be almost converged. All together, this system would benefit of longer simulation time. Nonetheless, each replicate has been run for 4 µ s, which after concatenation and removal of initial equilibration time resulted in a 19.25 µs long simulation of the system. Therefore, the usage of an enhanced sampling technique is probably more relevant. However, from the overall appearance of this system we do not expect drastically changed average values of the properties analyzed, with more sampling. Therefore this system is regarded as sufficiently sampled for allowing comparison to the other systems in this work.

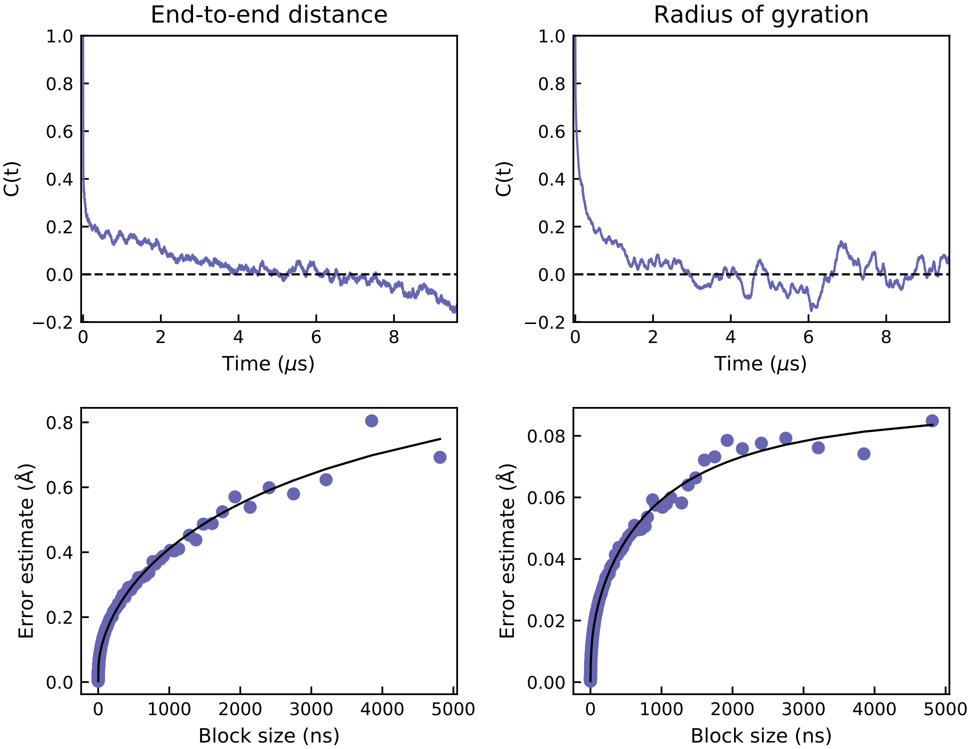


Figure S19: Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (left) and the radius of gyration (right) for the concatenated simulation of SN15p using CHARMM36m.

2 Effect of salt concentration in simulations

The simulations in this study were all performed using only ions to neutralize the net charge in the system, while the experiments were performed at an ionic strength of 150 mM. This section compares simulations without salt and with salt corresponding to a concentration of 150 mM for the phosphorylated peptide, using the CHARMM36m force field, to show that the salt-free simulations are still comparable with the experiments.

2.1 Convergence and sampling in the simulation of SN15p using CHARMM36m and 150 mM NaCl

The SN15p peptide simulated with CHARMM36m and 150 mM NaCl shows rather broad diversity in the energy landscape between different replicates, as seen in Figure S20, where especially replicate number two, three and five all have narrow, almost non-overlapping distributions in space. This suggest that especially replicate two and five samples a limited set of conformational space. The first and the fourth replicate show wider distributions with good agreement between them. The third replicate, and to some extent the second, appears to sample subspaces of the first and fourth replicate, while the fifth replicate samples a different conformational space. A visualization of the trajectories shows that in the second

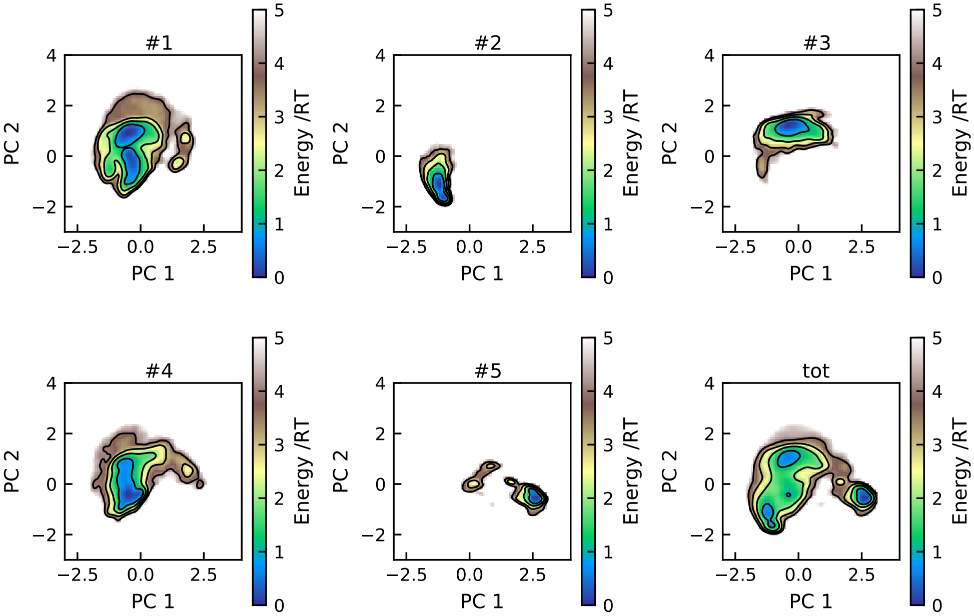


Figure S20: Energy landscapes for the five replicates and the concatenated trajectory of SN15p in CHARMM36m with 150 mM NaCl, using the two first principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 4$.

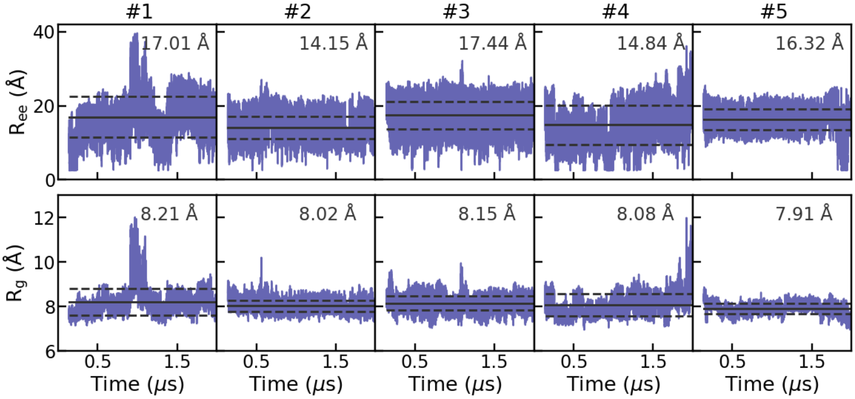


Figure S21: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of SN15p using CHARMM36m with 150 mM NaCl. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

replicate almost all conformations show hydrogen bonding between the arginines and one of the phosphoserines, while the other phosphoserine coordinates to lysine. In the fifth replicates all three arginines are hydrogen bonded to both phosphoserines. It appears that these conformations with a lot of hydrogen bonds (salt bridges) are so favourable that it can be problematic to break the interactions and change to another favourable conformation.

Like the salt-free simulation, the time evolutions of R_{ee} and R_g (Figure S21) show rather small spread, in agreement with a limited conformational ensemble. Only the first and the fourth replicate show time periods of drastically different R_{ee} or R_g. The drastic increase in R_{ee} and R_g in the first replicate around 0.9 μs is due to a conformational change where Arg13 is no longer hydrogen bonding to either phosphoserine, which allows for a more stretched out conformation. The same change is observed in the end of the fourth replicate. The density distributions of the end-to-end distance, shown in Figure S18a, are clearly bimodal, which is explained by the peptide having the ends close together or further away depending on whether Arg13 is hydrogen bonding to the phosphoserines or not. All replicates show distributions in the same range and only the peak height of the two main peaks differs,

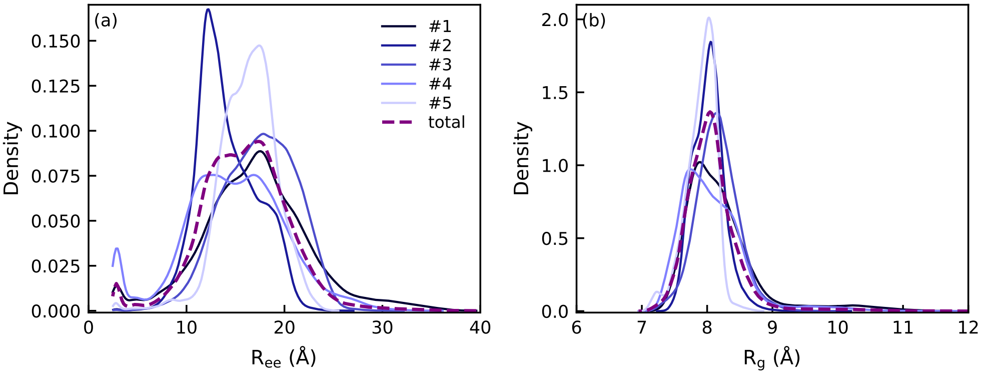


Figure S22: Density estimates of the end-to-end distance, $\mathrm{R_{ee}}$ (a) and the radius of gyration, $\mathrm{R_g}$ (b), for the five replicates in the simulation of SN15p using CHARMM36m with 150 mM NaCl, obtained from a Gaussian kernel estimator. The dashed purple line shows the distribution for the five replicates combined.

hence, the combined trajectory appears to give a rather good description of the system. The $\mathrm{R_g}$ density distributions (Figure S22b) are in better agreement than the $\mathrm{R_{ee}}$ density distributions, since they all are centered around almost the same value, only having slightly different widths.

The autocorrelation function of the total simulation, presented in Figure S23 is decreasing towards zero relatively fast, although later on showing some fluctuations around zero for both the end-to-end distance and radius of gyration. However, the error estimates from block averaging, shown in the same figure, converge nicely. It therefore appears that although the energy landscapes of some of the replicates showed little resemblance, the combined simulation is sampled well enough to allow comparison with the corresponding system without salt.

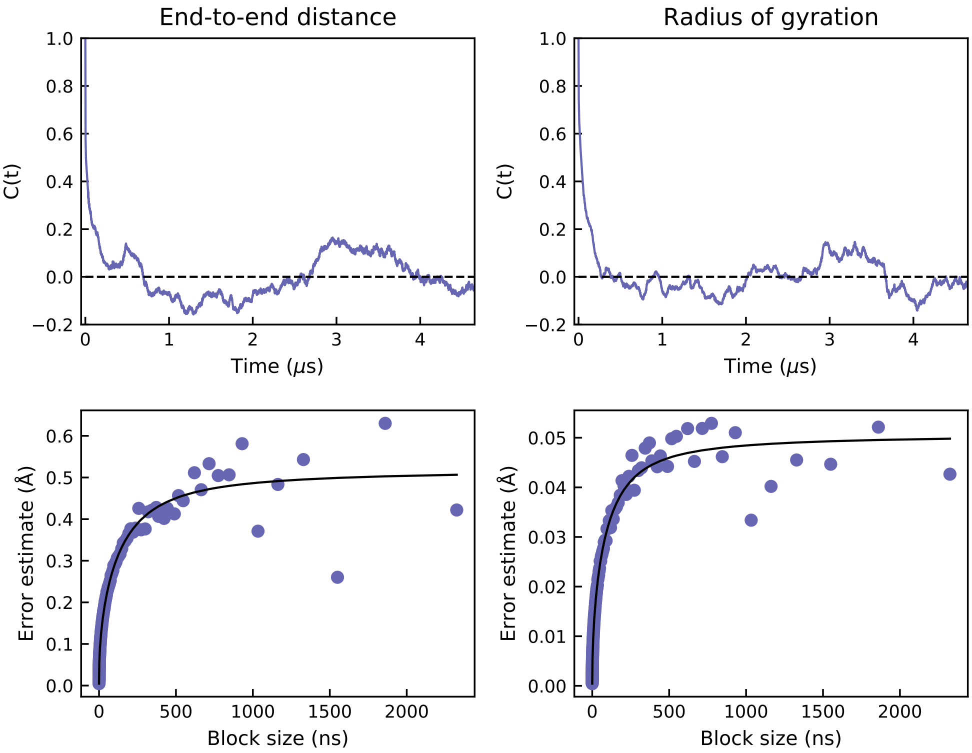


Figure S23: Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (left) and the radius of gyration (right) for the concatenated simulation of SN15p using CHARMM36m with 150 mM NaCl.

2.2 SN15p with CHARMM36m, with and without 150 mM NaCl

In Figure S24 the density estimates of R_{ee} and R_g, and the shape in the form of a dimensionless Krakty plot, is compared between having no salt except for counterions and having Na and Cl ions corresponding to a salt concentration of 150 mM. It appears that the added salt has no effect on these properties.

Regarding the secondary structure, the Ramachandran plots in Figure S25 are close to identical, with the only difference being a slightly more populated turn region and alpha-helical region in the presence of 150 mM NaCl. The higher turn content is supported by the DSSP analysis, see Figure S26. The simulation without salt instead shows a higher bend

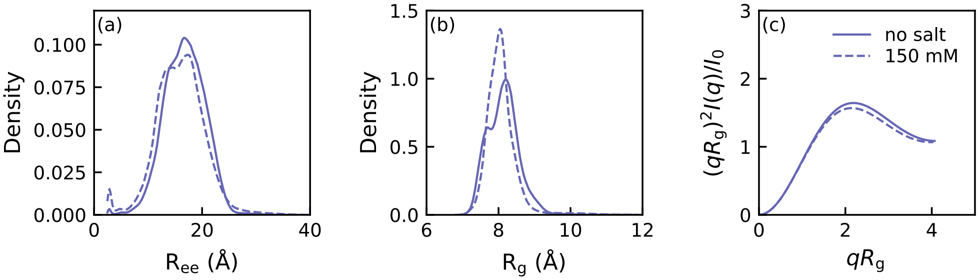


Figure S24: Density estimates of the end-to-end distance, R_{ee} (a) and the radius of gyration, R_g (b), and a dimensionless Kratky plot of SN15p simulated using CHARMM36m with and without 150 mM NaCl.

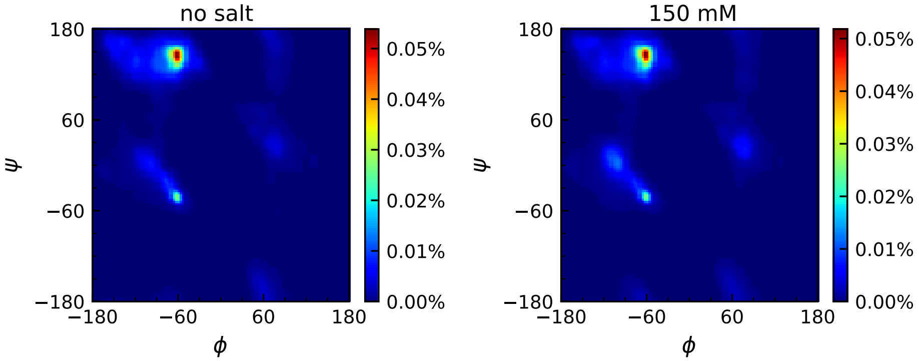


Figure S25: Ramachandran plots of SN15p, simulated with CHARMM36m without (left) and with (right) 150 mM NaCl. The color scale shows the population density.

content, while in both cases the content of β-sheets, β-bridges, and helices are very low.

The area where the salt is expected to have the largest effect is in the contacts between phosphoserine and arginine. The contact map, see Figure S27, shows great similarity between having additional salt or not. On average, the probability of contact between arginine and phosphoserine is reduced by approximately 10 percentage points, although the contact between pSer3 and Arg9 is increased. Also the contact between phosphoserine and lysine is increased. Since both of the simulations would benefit of more sampling before extracting exact numbers with high certainty, some variation is expected. The important part is that

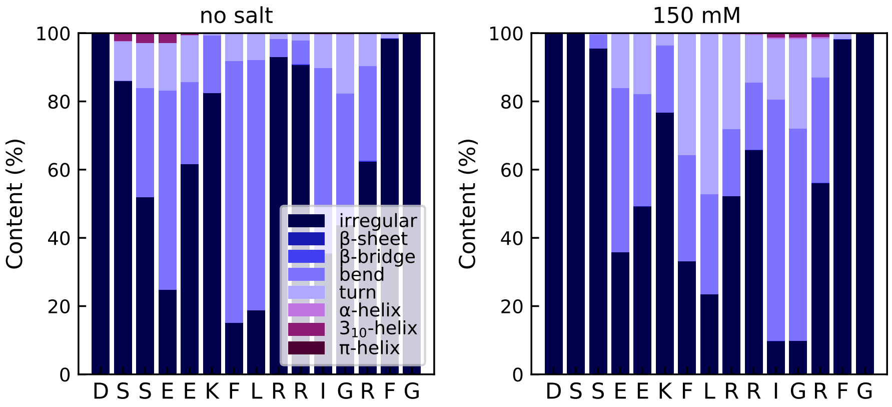


Figure S26: Stacked bar chart of the secondary structure content determined by DSSP of each amino acid in the SN15p peptide, simulated with CHARMM36m without (left) and with (right) 150 mM NaCl. The legend applies to both panels.

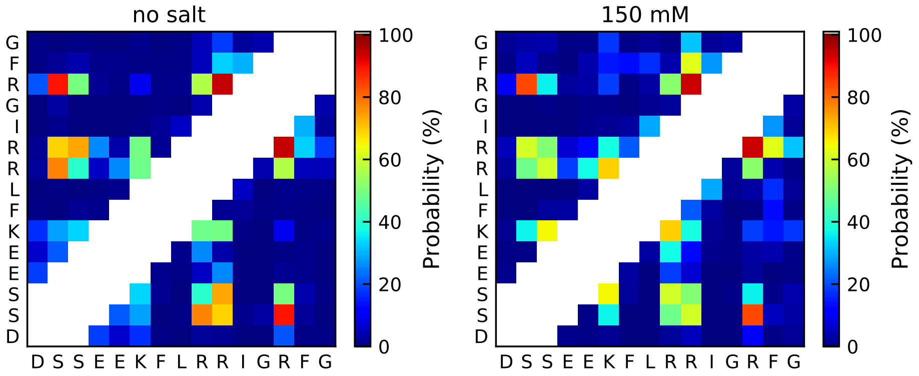


Figure S27: Contact map showing the probability of atoms in different residues being closer than 4 Å in SN15p simulated with CHARMM36m without (left column) and with (right column) 150 mM NaCl. The two closest residues on each side as well as the residue itself are excluded from the analysis and therefore shown in white.

even in the presence of 150 mM NaCl, the interactions between arginine and phosphoserine are highly conserved. Also the total number of intrapeptide hydrogen bonds are highly similar in the two different conditions, see Figure S28. However, there is a higher probability for having a few more helical hydrogen bonds in a conformation in the presence of 150 mM NaCl, which is balanced out by the most probable number of more distant hydrogen bonds

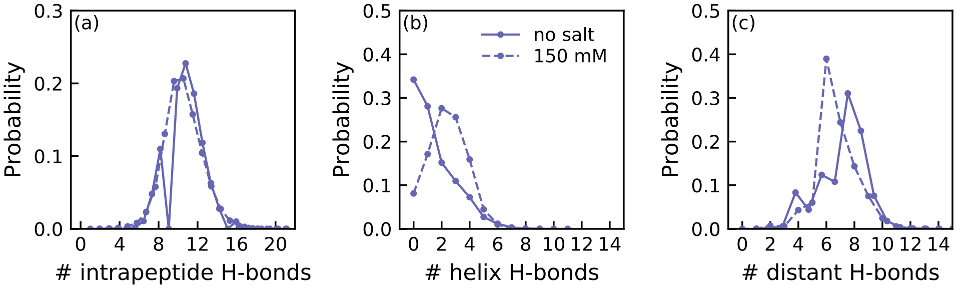


Figure S28: Probability distribution of the number of intrapeptide hydrogen bonds in total (a), associated with helices, i.e, between residue n and n+i, where i=3, 4, 5 (b), and between residue n and n+i, where i≥ 6 (c). The legend applies to all panels.

being lowered by one. This is in agreement with the lysine–phosphoserine contact being more probable, which falls in the helical category due to the residues separation, and the probability of arginine–phosphoserine contacts being lower.

To summarize, neither the overall shape or size of the peptide is affected by the ionic strength. There are some differences in the dominating secondary structure being irregular, bend or turn for individual amino acids, while both simulations agree on the structural content being low. The probability of arginine–phosphoserine contacts is reduced by approximately 10 percentage points in the presence of salt, although still being highly conserved. In addition, the probability of lysine–phosphoserine contact is instead increased, suggesting that the electrostatic interactions and hydrogen bonds within the peptide is still of high importance in the presence of 150 mM NaCl. Altogether it can be concluded that the effect of salt on the simulations of this system is minor, especially on the global structural which the experimental data describes. Hence, it is valid to compare the simulations with the experimental data, despite them lacking the ionic strength of 150 mM which was used in experiments.

3 Small angle X-ray scattering data

The full concentration series measured are shown in Figure S29, with the determined radius of gyration and molecular weight presented in Table S1 and S2.

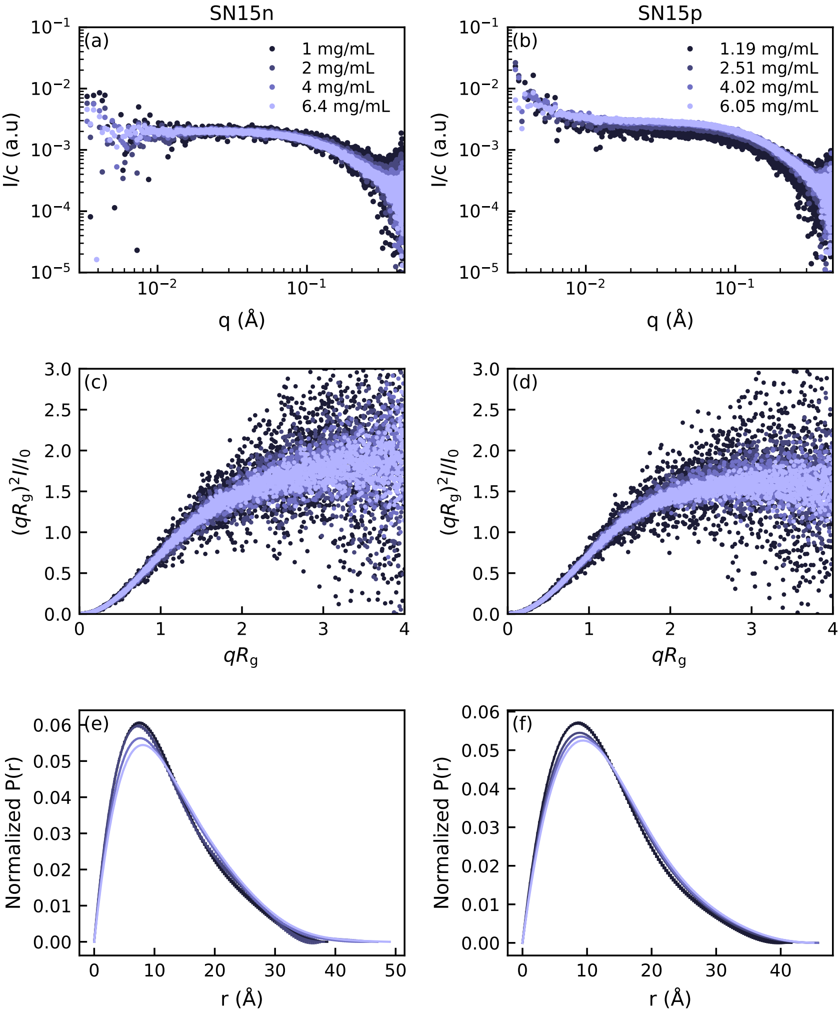


Figure S29: SAXS data for SN15n (a,c,e) and SN15p (b,d,f) obtained at 20 mM Tris, 150 mM NaCl, pH 7.5, and 20 °C. Scattering intensity curve (a,b), dimensionless Kratky plot (c,d) and pair distance distribution function, P(r) (e,f). The legends in the upper panels apply to the full column.

Table S1: Radius of gyration and molecular weight for the SN15n samples. The molecular weight is calculated from the I_0 determined from P(r).

c (mg/mL)	$R_{g,P(r)}$ (Å)	$R_{g,Guinier}$ (Å)	M_w (kDa)
1	9.9 ± 0.2	9.1 ± 0.3	2.09
2	10.0 ± 0.2	9.6 ± 0.1	2.03
4	10.5 ± 0.2	9.9 ± 0.1	2.09
6.4	10.9 ± 0.1	10.2 ± 0.2	2.05

Table S2: Radius of gyration and molecular weight for the SN15p samples. The molecular weight is calculated from the I_0 determined from P(r).

c (mg/mL)	$R_{g,P(r)}$ (Å)	$R_{g,Guinier}$ (Å)	M_w (kDa)
1.19	10.5 ± 0.2	9.6 ± 0.6	2.06
2.51	10.9 ± 0.1	10.1 ± 0.8	2.54
4.02	11.1 ± 0.1	10.5 ± 0.8	2.92
6.05	11.3 ± 0.1	10.7 ± 0.1	3.01

4 Analysis of circular dichroism spectra with BeStSel

To obtain the partition of secondary structure in the peptides, the experimental data were analyzed with BeStSel. Figure S30 shows the fit in comparison to the experimental data.

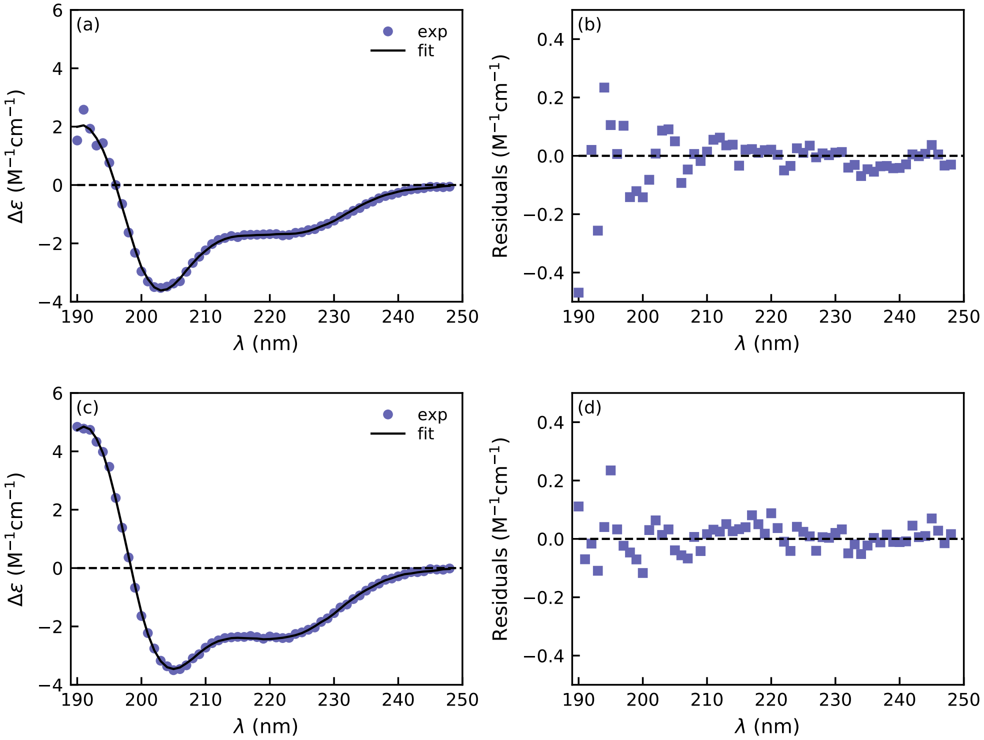


Figure S30: Experimental CD data and the fit from BeStSel expressed as the molar differential extinction coefficient, $\Delta\varepsilon$, for SN15n (a), and SN15p (c), with corresponding residuals in (b) and (d), respectively.

Paper IV

International Journal of
Molecular Sciences

MDPI

Article

Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison

Ellen Rieloff [1] and Marie Skepö [1,2,*]

[1] Division of Theoretical Chemistry, Lund University, POB 124, SE-221 00 Lund, Sweden; ellen.rieloff@teokem.lu.se

[2] LINXS—Lund Institute of Advanced Neutron and X-ray Science, Scheelevägen 19, SE-223 70 Lund, Sweden

* Correspondence: marie.skepo@teokem.lu.se

Abstract: Phosphorylation is a common post-translational modification among intrinsically disordered proteins and regions, which helps regulate function by changing the protein conformations, dynamics, and interactions with binding partners. To fully comprehend the effects of phosphorylation, computer simulations are a helpful tool, although they are dependent on the accuracy of the force field used. Here, we compared the conformational ensembles produced by Amber ff99SB-ILDN+TIP4P-D and CHARMM36m, for four phosphorylated disordered peptides ranging in length from 14–43 residues. CHARMM36m consistently produced more compact conformations with a higher content of bends, mainly due to more stable salt bridges. Based on comparisons with experimental size estimates for the shortest and longest peptide, CHARMM36m appeared to overestimate the compactness. The difference between the force fields was largest for the peptide showing the greatest separation between positively charged and phosphorylated residues, in line with the importance of charge distribution. For this peptide, the conformational ensemble did not change significantly upon increasing the ionic strength from 0 mM to 150 mM, despite a reduction of the salt-bridging probability in the CHARMM36m simulations, implying that salt concentration has negligible effects in this study.

Keywords: intrinsically disordered proteins; phosphorylation; force fields

check for updates

Citation: Rieloff, E.; Skepö, M. Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison. *Int. J. Mol. Sci.* **2021**, *1*, 1. https://doi.org/

Academic Editor:

Received:
Accepted:
Published:

1. Introduction

Intrinsically disordered proteins (IDPs) are characterized by a lack of a tertiary structure under physiological conditions [1,2], which means that they are better described by an ensemble of different conformations than a single structure. This is reflected in their free energy landscapes, which normally are rather flat without a deep energy minimum as for globular proteins [3]. The flattened energy landscape makes IDPs very sensitive to changes in the environment and post-translational modifications (PTMs) of the sequence. A common type of reversible PTM is phosphorylation, which introduces extra negative charges and the possibility of forming hydrogen bonds and salt bridges [4]. Phosphorylation is commonly employed by cells as a regulatory mechanism, as it can change both the conformational ensemble and the dynamics, as well as the interaction with a binding partner, and therefore affect function. The functional implications of phosphorylation can be drastic, such as for the disordered neuroprotein tau, for which hyperphosphorylation has been related to amyloid fibril formation in Alzheimer's disease [5]. In proteins such as statherin and caseins, the phosphorylated residues are essential for their ability to bind to the tooth surface [6,7] or sequester calcium [8].

Experimental techniques such as small-angle X-ray scattering (SAXS) and fluorescence resonance energy transfer (FRET) have been used to provide information on global conformational changes upon phosphorylation of intrinsically disordered proteins or regions, while circular dichroism spectroscopy and nuclear magnetic resonance (NMR) have detected changes in secondary structure or other local arrangements such as salt bridges [9–14]. However, due to the vast conformational ensembles possessed by IDPs, computer simulations are often a useful complement to obtain more detailed information, though this requires accurate models and force fields. We have previously shown that a coarse-grained "one bead per residue model" has proven to accurately predict average radius of gyration (R_g) and scattering curves for various IDPs, including statherin, although producing overly compact conformations of other more phosphorylated IDPs [15]. The two-site UNRES model has recently been extended with parameters for phosphorylated residues [16] and applied to study phosphorylation-induced folding of an IDP [17]. Although coarse-grained models are more computationally efficient and generally easier to interpret than atomistic models, they can lack in detail. In atomistic modelling, there is continuous development of force fields and water models towards more accurately describing IDPs, and some important adjustment have been the refinement of the backbone dihedral angles and balancing the water–protein and protein–protein interactions; see for example the following reviews and references within [18,19]. However, we recently showed that while the commonly used force fields CHARMM36m and Amber ff99SB-ILDN+TIP4P-D accurately captured the global dimensions of the 15-residue-long N-terminal fragment of Statherin in the nonphosphorylated state, it overestimated the compactness in the phosphorylated state [20]. More recently, overcompaction was also observed for two approximately 80-residue-long phosphorylated IDPs in several force fields, where it was suggested to depend on an overestimation of charge–charge interactions [21], in line with an overstabilization of salt bridges in standard force fields [22]. In this study, we made a further comparison of the two aforementioned force fields, by applying them to four phosphorylated peptides, namely two different fragments from tau, specifically residues 173-183 (Tau1) and 225-246 (Tau2), the first 25 amino acids in the milk protein β-casein (bCPP) and the saliva protein statherin (Stath). For all peptides, CHARMM36m was shown to sample more compact conformations than Amber ff99SB-ILDN+TIP4P-D, associated with a much higher probability for salt bridges. The effect was more pronounced in sequences with large separation between phosphorylated residues and positively charged residues, showing the importance of charge distribution. In bCPP, which showed the largest differences between the force fields, the addition of 150 mM NaCl did not change the average size estimates and shape significantly, despite a significant reduction of salt bridge occurrence in CHARMM36m. This implies that salt bridges are still of importance at 150 mM salt and that we can ignore the effects of salt concentration in this study.

2. Results and Discussion

Four phosphorylated peptides, shown in Table 1, were simulated at physiological pH using two different force fields: Amber ff99SB-ILDN [23] with the TIP4P-D [24] water model and parameters for the phosphorylated residues from Homeyer et al. [25] and Steinbrecher et al. [26] (A99) and CHARMM36m [27] with the CHARMM-modified TIP3P water model [28] (C36). The peptides were chosen based on availability of experimental data to compare with and size considering the computational expense.

Table 1. Full name and sequence of the peptides included in this study. Positively charged residues are marked in blue, negatively charged in red, and phosphorylated residues highlighted with yellow. Note that Tau1 includes three additional residues in accordance with [11], to allow for experimental comparison.

Name	Protein	Sequence
Tau1	Tau$_{173\text{-}183}$	CAKTPPAPKTPPAW
Tau2	Tau$_{225\text{-}246}$	KVAVVRTPPKSPSSAKSRLQTA
bCPP	β-casein$_{1\text{-}25}$	RELEELNVPGEIVESLSSSEESITR
Stath	Statherin	DSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF

2.1. Size and Shape

For all four peptides, the two force fields produced different conformational ensembles, as seen by the distributions of the R_g and the end-to-end distance (R_{ee}) in Figure 1. The C36 distributions were narrower and centered on values lower than the A99 distributions. For Tau2 and bCPP, the R_g distribution had a sharp peak at low values. From the average R_g and R_{ee} presented in Table 2, it is clear that Tau1 showed the smallest differences between the force fields, while bCPP showed the largest differences. The discrepancy was larger for R_{ee} than R_g. For Tau1, Chin et al. [11] determined the average R_{ee} to be ∼3.17 nm, based on FRET. To obtain an R_{ee} distance distribution from the FRET data they assumed a semi-flexible polymer model, and the resulting distribution was skewed towards longer distances, with the peak value located at 3.64 nm (Figure 4A in ref. [11]). Comparing A99 and C36 to the experimental average, A99 overestimated it approximately as much as C36 underestimated it. However, the skewed shape and peak position at 3.64 nm produced in A99 was in better experimental agreement than C36, since the distribution in C36 was more symmetrical with multiple peaks and had the main peak located at 3.03 nm.

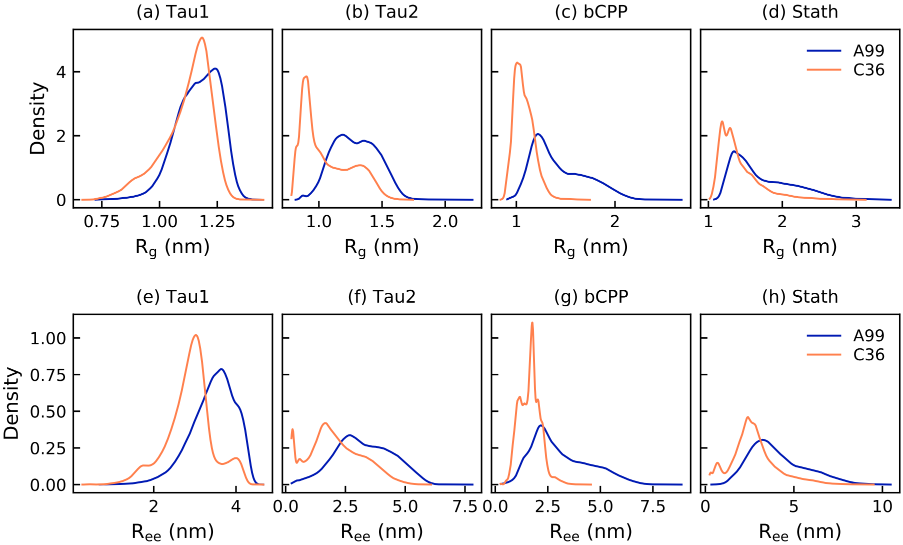


Figure 1. Distribution of the radius of gyration (**top** row) and the end-to-end distance (**bottom** row) of Tau1, Tau2, bCPP, and Stath simulated with Amber ff99SB-ILDN (A99) and CHARMM36m (C36). The legend applies to all panels.

Table 2. Average radius of gyration and end-to-end distance of the peptides simulated with Amber ff99-SB-ILDN (A99) and CHARMM36m (C36). The difference between the force fields is expressed in relation to A99.

Peptide	Radius of Gyration (nm)			End-to-End Distance (nm)		
	A99	C36	Difference (%)	A99	C36	Difference (%)
Tau1	1.17 ± 0.01	1.12 ± 0.01	4	3.44 ± 0.04	2.88 ± 0.07	16
Tau2	1.29 ± 0.03	1.06 ± 0.10	18	3.27 ± 0.17	2.10 ± 0.32	36
bCPP	1.43 ± 0.03	1.08 ± 0.02	24	3.09 ± 0.15	1.65 ± 0.10	47
Stath	1.73 ± 0.09	1.41 ± 0.04	18	4.05 ± 0.17	2.74 ± 0.20	32

For Stath, earlier published SAXS data [15] provided an R_g of 1.93 ± 0.2 nm; hence, R_g was 10% smaller in A99 and 27% smaller in C36. Since R_g determined from SAXS includes a hydration shell, it was expected that R_g calculated from simulations would be slightly smaller, although not to that extent. Since it is not straightforward which contrast to use for the hydration shell in the calculations of scattering curves for IDPs [29], in Supplementary Figure S1 and Table S2, we compared the curves calculated using different contrasts of the hydration shell to the experimental curve for Stath. While the highest contrast used (0.03 $e/Å^3$) yielded the best agreement with the scattering curve, it provided the worst agreement with the Kratky plot. Henriques et al. [29] showed that the optimal contrast for IDPs was often between 0.01 $e/Å^3$ and 0.02 $e/Å^3$, although varying with both force field and protein. The optimal values for A99 and C36 were suggested to be around 0.0075 $e/Å^3$ and 0.02 $e/Å^3$, respectively. While the suggested optimal value gave reasonable agreement with the experimental form factor for A99, this was not the case for C36. For C36, all contrasts > 0 clearly showed larger compaction than the experimental Kratky plot.

Even without experimental scattering curves to compare to, the dimensionless Kratky plot, presented in Figure 2, is a good way of comparing the average shape of the peptides in the two different force fields. The short peptide Tau1 exhibited a more extended shape than the other three peptides, which in A99 were shown to have more of the typical IDP behavior, resembling a Gaussian chain. For all four peptides, the Kratky plot produced in C36 had a lower slope, and for the three longest peptides, the curve started to move towards the bell-shaped curve typical of globular proteins. Hence, this implies that C36 sampled more compact or well-defined conformations than A99, in accordance with the R_g and R_{ee} distributions. Notice also that the Kratky plot of Stath in A99 was in excellent agreement with the experimental data, while the curve corresponding to C36 fell below, as shown in Figure 2d.

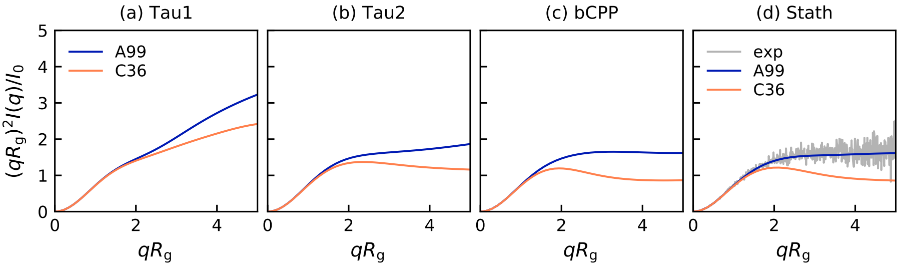


Figure 2. Dimensionless Kratky plot from simulations with Amber ff99SB-ILDN and CHARMM36m for (**a**) Tau1, (**b**) Tau2, (**c**) bCPP, and (**d**) Stath. In Panel (d), experimental data from Cragnell et al. [15] are included for comparison. The legend in Panel (a) is applicable to all panels.

2.2. Salt Bridges and Secondary Structure

Since our previous study [20] suggested that overstabilized salt bridges are the reason why C36 produces more compact conformations than A99, we calculated the occupancy of the possible salt bridge interactions involving the phosphorylated residues. Figure 3 indeed shows that salt bridges were formed much more in C36 than A99, for all the peptides. In Tau2 and bCPP, the strong salt bridges in C36 restricted the conformational ensemble, which explains the smaller and narrower distributions of R_g and R_{ee}. In bCPP, the salt-bridging residues were well separated in the sequence, therefore having a larger effect on the R_g and R_{ee} distributions. In Tau1, the salt bridge interactions almost exclusively appeared between the adjacent residues and between pT175 and the N-terminal.

For Tau2, there is experimental evidence of the following salt bridges, detected by NMR experiments: pT231–R230, pS237–K240, and pS238–R242 [12]. pT231–R230 and pS238–R242 are indeed two of the most often occurring salt bridges in A99, while pS237–R242 is more common than pS237–K240. Several other salt bridges are also as frequently present as pS237–K240. In C36, pT231–R230 is the most occurring salt bridge, but both pS327–R242 and pS235–K234 are more probable than pS237–K240. Hence, while both force fields captured the experimentally established salt bridges, they also suggested other salt bridges to be present and some of them to be more common than the experimentally established ones.

Advancing to the secondary structure, Figure 4 shows that the peptides were mainly irregular, although Tau1 contained much of the polyproline type II (PPII) structure as well. In fact, all peptides contained a significant amount of PPII, as well as a significant content of bends. The content of the helical structure (α- and 3_{10}-helix) and β-strands was low in all peptides. Tau1 exhibited the largest differences between the force fields, where A99 produced 16 percentage points more of the PPII structure than C36, which instead contained a more irregular structure. For the other peptides, the differences were smaller. Overall, the peptides only had one significant difference in common, which was a higher content of bends in C36 than A99. Inspecting the content along the sequence, it was evident that it was mostly the same parts of the peptide that were enriched in a certain type of structure in both force fields (see Supplementary Figure S3). However, in C36, the helical content was completely missing from the first ten residues of Stath, which is concerning since the N-terminal region has been shown to possess helical propensity in water, although being mainly disordered [6,30]. Another striking difference between the force fields for Stath is that some residues centered on residues Y21 and Y41 occasionally formed a β-sheet or β-bridge in C36, but not in A99. Notice also that for Tau2, the bend propensity at residues V228–V229 was much higher in C36 than in A99. Since these residues were located right between K225 and pT231, which in C36 formed a stable salt bridge, this suggested that the bend was formed as a result of the salt bridge. Furthermore, for Tau2, NMR data have suggested approximately 40% α-helical propensity in region A15-R18 [12]. Both A99 and C36 sampled the helical structure in this region, however, to a lower extent than what the experimental data suggested.

2.3. Energy Landscapes

The differences between the force fields in this study is well summarized by the energy landscapes in Figures 5–8. Tau2, bCPP, and Stath all showed a narrower energy landscape in C36, in line with a more restricted conformational ensemble. Tau1, which is rather short and stiff, actually gained a larger conformational landscape in C36, due to sampling more bent conformations in addition to being more stretched out as in A99; see Figure 5. Notice also that in C36, the global minimum, which was the most populated, contained conformations that were not entirely stretched out. Instead, the N-terminal end was folded over, such that a salt bridge was formed between pT175 and the positively charged N-terminus.

Although the energy landscapes of Tau2 in A99 and C36 were located in almost the same area, the energy levels differed; see Figure 6. The most populated basin in the C36 simulation was a deep and narrow minimum, while the A99 simulation had a larger area of energy $\leq$1RT, containing several basins, more typical of IDPs. The salt bridges creating more compact conformations were evident in the C36 conformations, while the A99 conformations were more stretched out with fewer salt bridges. Notice that the phosphorylated residues in C36 had a tendency to interact with several positively charged residues simultaneously. In both force fields, a basin minimum with a helical region starting with pS237 and pS238 was found, in line with the secondary structure analysis.

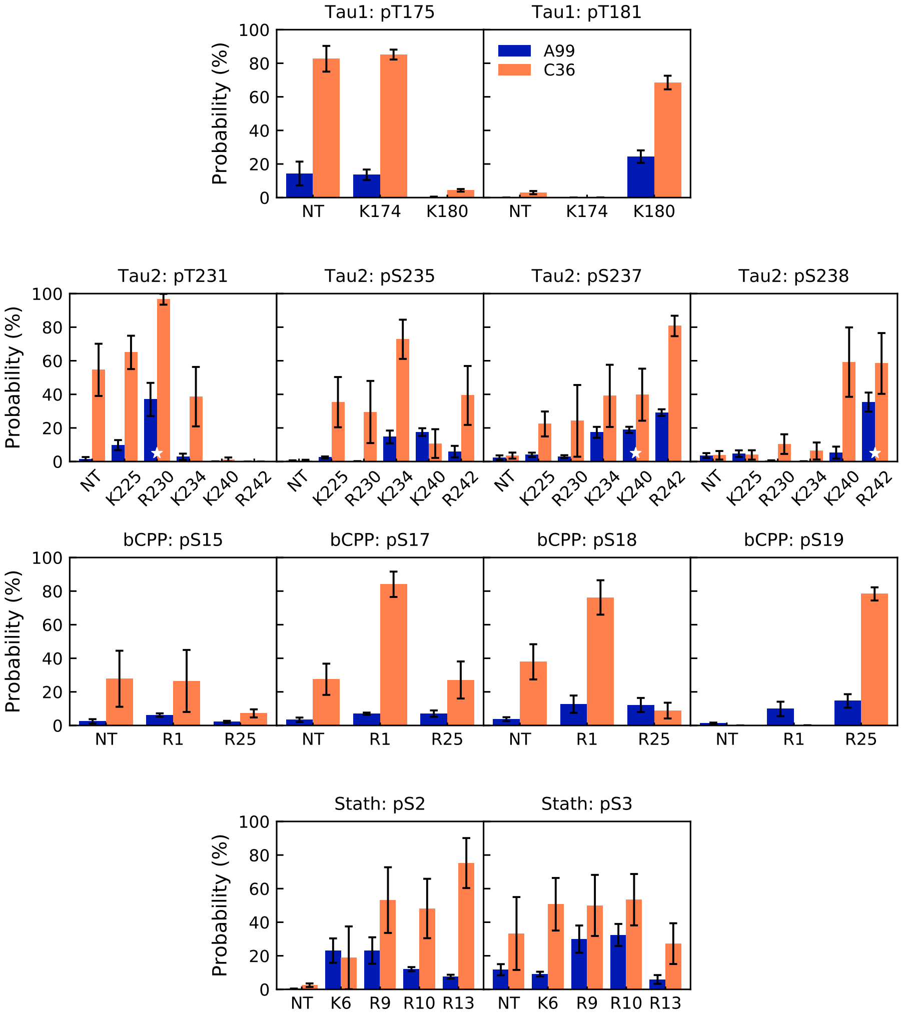


Figure 3. Probability of possible salt bridge interactions for the phosphorylated residues with the N-terminus (NT) and positively charged residues in Tau1 (**first** row), Tau2 (**second** row), bCPP (**third** row), and Stath (**last** row). For Tau2, experimentally established salt bridges [12] are marked with a white star. Error bars correspond to errors calculated by block averaging.

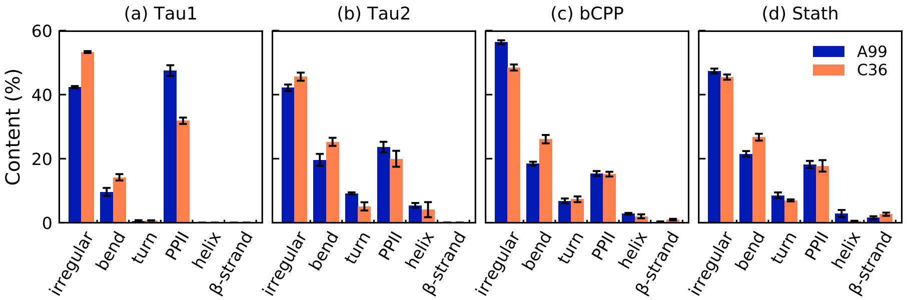


Figure 4. Average content of different types of the secondary structure in (**a**) Tau1, (**b**) Tau2, (**c**) bCPP, and (**d**) Stath simulated with Amber ff99SB-ILDN (A99) and CHARMM36m (C36). The legend applies to all panels. The helix includes the α- 3_{10}- and a negligible content of the π-helix, while the β-strand also includes β-bridge. Error bars correspond to errors calculated by block averaging.

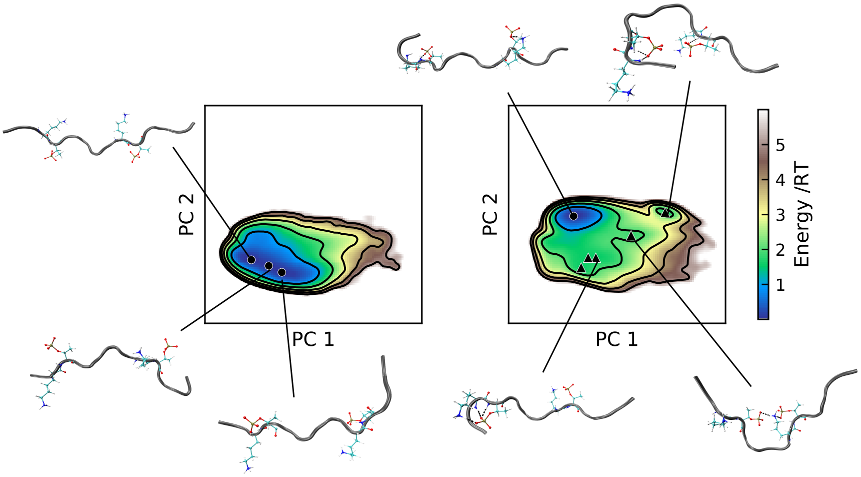


Figure 5. Energy landscapes and conformations in selected minima of Tau1. (**Left**) A99; (**right**) C36. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both force fields, such that they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq \mathrm{RT} \leq 5$, and the minimum of each basin is represented by a marker: ●: energy $\leq 1RT$, ▲: $\leq 2RT$. In the conformations, the phosphorylated and positively charged residues are shown explicitly. Dashed black lines represent hydrogen bonds. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil) and cyan is turns. The N-terminus of each conformation is the leftmost end.

For bCPP, there was indeed many more elongated conformations in the A99 simulation (see Figure 7), and it is clear that what caused the more compact conformations in C36 was the salt bridges between the phosphorylated serines and the arginines. In C36, all depicted conformations contained at least one salt bridge between phosphoserine and arginine, while this was much rarer in A99, explaining why the energy landscapes

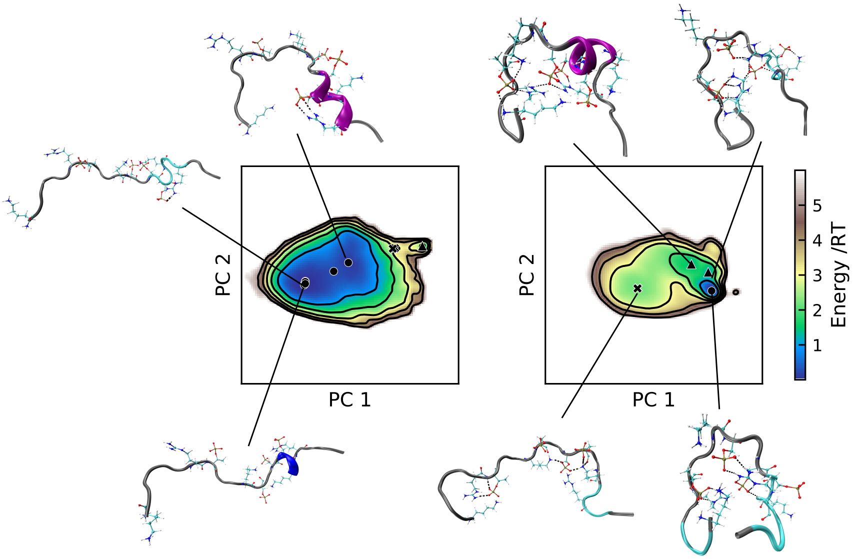


Figure 6. Energy landscapes and conformations in selected minima of Tau2. (**Left**) A99; (**right**) C36. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both force fields, such that they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq \mathrm{RT} \leq 5$, and the minimum of each basin is represented by a marker: ●: energy $\leq 1RT$, ▲: $\leq$2RT, ✖: $\leq$3RT. In the conformations, the phosphorylated and positively charged residues are shown explicitly. Dashed black lines represent hydrogen bonds. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil), cyan is turns, magenta is the α-helix, and blue is the 3_{10}-helix. The N-terminus of each conformation is the leftmost end.

looked so different. Regarding Stath, comparing the conformations in Figure 8, there were two striking differences. First, there was a higher presence of salt bridges between phosphoserine and positively charged residues in C36, keeping the N-terminal end in a more bent conformation. Secondly, in C36, the β-strand and β-bridge formation between the middle region and C-terminal region detected in Supplementary Figure S3 contributed to making the conformations more compact compared to A99.

2.4. Effect of Salt Concentration

Since the salt bridges formed between phosphorylated and positively charged residues were shown to influence the conformational ensemble, it is of importance to also consider the effect of the screening of the electrostatic interactions. Here, we focused on bCPP, which due to showing the largest differences between force fields and having the highest fraction of charged residues in combination with the largest charge separation (see Supplementary Table S1), was expected to show the largest response to ionic strength. Figure 9 shows that in C36, four of the salt bridges were dramatically reduced upon the addition of 150 mM NaCl; however, the probability of two other salt bridges increased, whereas in A99, only one salt bridge was significantly reduced. At 150 mM salt, the salt-bridging probability was more comparable between A99 and C36, although overall still higher in C36. Supplementary Figure S3 shows the changes in the contact map upon the addition of 150 mM NaCl for bCPP simulated in A99 and C36. For A99, we clearly saw that the preference for the N-terminal end to be in contact with the phosphorylated and negatively charged region

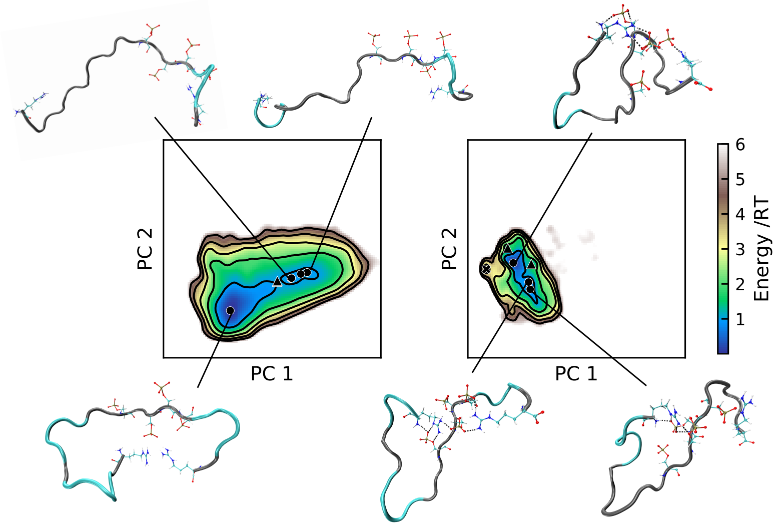


Figure 7. Energy landscapes and conformations in selected minima of bCPP. (**Left**) A99; (**right**) C36. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both force fields, such that they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq \text{RT} \leq 5$, and the minimum of each basin is represented by a marker: ●: energy ≤ 1RT, ▲: ≤2RT, ✖: ≤3RT. In the conformations, the phosphorylated and positively charged residues are shown explicitly. Dashed black lines represent hydrogen bonds. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil) and cyan is turns. The N-terminus of each conformation is the leftmost end.

(residues 14–21) diminished. In C36, the strongly conserved R1–pS17 and R1–pS18 contacts were greatly decreased, while the contact of R1 with surrounding residues in the negatively charged region was increased. Hence, this suggested an increased mobility, while still maintaining contact with the negatively charged region. In C36, the cross-diagonal lines also signalized a decrease of the β-sheet; however, the content was relatively low from the beginning.

By comparing the energy landscapes in Figure 10, it is clear that screening of the electrostatic interactions indeed broadened the conformational ensemble, but mainly in C36, which also showed the largest change in salt bridge probability. In C36, the addition of 150 mM NaCl led to the exploration of more stretched out conformations; however, more compact conformations still clearly dominated. A99 also showed an increased probability of visiting more stretched out conformations after the addition of 150 mM NaCl. This shift in the conformational ensemble was also observed in the distributions of R_g and R_{ee} shown in Supplementary Figure S4. However, the changes were actually rather small, such that the average values were indistinguishable. Upon the addition of salt, the R_g changed from 1.43 ± 0.03 nm to 1.45 ± 0.03 nm for A99 and from 1.08 ± 0.02 nm to 1.08 ± 0.03 nm for C36. The changes in R_{ee} were from 3.09 ± 0.15 nm to 3.37 ± 0.13 nm and from 1.65 ± 0.10 nm to 1.67 ± 0.10 nm, respectively. The effect of salt on the calculated scattering curves was also so small that it could be deemed negligible; see Supplementary Figure S5.

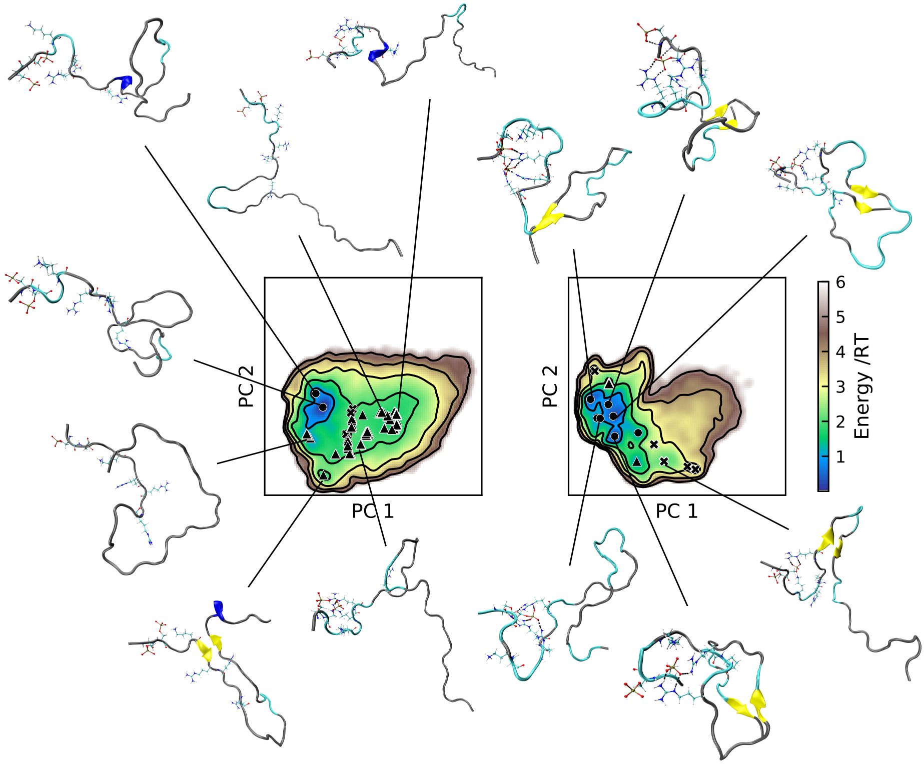


Figure 8. Energy landscapes and conformations in selected minima of Stath. (**Left**) A99; (**right**) C36. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both force fields, such that they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$, and the minimum of each basin is represented by a marker: ●: energy $\leq$ 1RT, ▲: $\leq$2RT, ✖: $\leq$3RT. In the conformations, the phosphorylated and positively charged residues are shown explicitly. Dashed black lines represent hydrogen bonds. The peptide conformations are color-coded according to the secondary structure determination in VMD, where silver is irregular (coil), cyan is turns, blue is the 3_{10}-helix, yellow is the β-sheet, and tan is the β-bridge. The N-terminus of each conformation is the leftmost/topmost end.

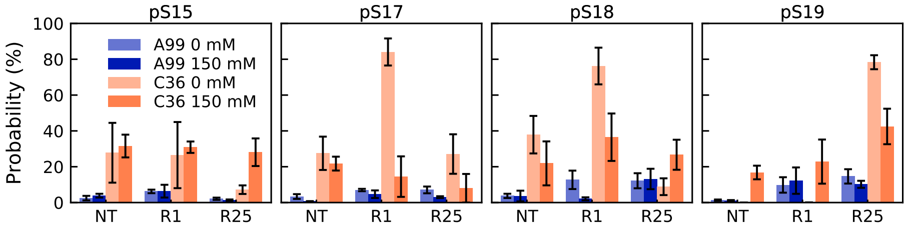


Figure 9. Probability of possible salt bridge interactions for the phosphorylated residues with the N-terminus (NT) and positively charged residues in bCPP, simulated with the two different force fields in the presence of 0 mM or 150 mM NaCl. Error bars corresponds to errors calculated by block averaging.

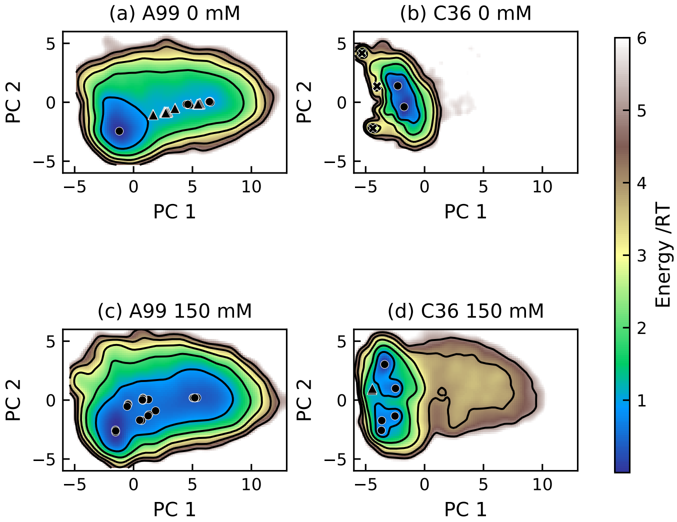


Figure 10. Energy landscapes of bCPP simulated with the two force fields Amber ff99SB-ILDN (A99) and CHARMM36m (C36) in the presence of 0 mM or 150 mM NaCl.

3. Conclusions

C36 produced more compact conformations of all four peptides, which indeed was expected to be caused mainly by salt bridge stability. In Tau1, the salt bridges pT175–K174 and pT181-K180 were formed without much effect on the overall conformation; however, an additional salt bridge between the N-terminus and pT175 decreased R_{ee} and R_g in C36. In Stath, the salt bridges contributed to the discrepancy by restricting the conformation of the first 15 residues, in the same way as previously shown for that fragment studied alone [20]. However, also the β-bridge and β-strand formation between the middle and C-terminal region were shown to contribute to more compact conformations. While C36 produced good results of nonphosphorylated short IDPs, it has been shown to underestimate the size of larger IDPs (>60 residues) [32,33]. Since Stath was 43 residues long, and thus the longest peptide included in this study, it is reasonable to believe that other effects also play a role. That bCPP showed the largest difference between the force fields and Tau1 the smallest implies that the separation between the phosphorylated and positively charged residues controls how much the conformational ensemble is influenced by stable salt bridges. This is in accordance with the importance of considering the level of charge separation for predicting the conformational ensemble of IDPs with a high fraction of charges [31].

When comparing to experimental data, it is important to consider the effect of salt, since most experiments are performed in the presence of buffer and additional salt. In bCPP, the addition of 150 mM NaCl was shown to dramatically reduce the probability of some of the salt bridges in C36, whereas the probability of other salt bridges actually increased. In A99, only one salt bridge was significantly reduced, which suggests that salt bridges still are of importance at 150 mM NaCl. Considering the changes in salt bridge probability for bCPP with salt concentration, it is plausible that the discrepancies between the simulations and experimental reference for Tau2 were caused by nonmatching ionic strength, since the experiments were performed with 50 mM phosphate buffer. At the same time, it can be hard to discern the salt bridges involving close-by residues experimentally, such as for pS237, pS238, K240, and R242.

Despite significant differences in the salt-bridging probability in C36, the effect of salt concentration on the global conformational level, such as R_g and R_{ee}, was small enough to be negligible for both force fields. In fact, the calculated form factor was indistinguishable, implying that comparing simulations performed without salt with experimental SAXS data collected at 150 mM NaCl indeed can be valid. Since bCPP is the peptide for which we expected the largest effects of salt concentration, this further strengthens the comparison with SAXS data for Stath collected at 150 mM NaCl, which showed that A99 was in good agreement, while C36 overestimated the level of compaction. Although the effects of ionic strength seem negligible in this study, this is generally not the case. For example, Jin and Gräter needed 350 mM of salt in simulations with A99 to reach experimental agreement for IDPs that are approximately 80 residues long [21], which suggested that also A99 overestimate the strength of salt bridges. Here, both Tau1 and Stath were compared to experimental size estimates, and only C36 was with certainty shown to underestimate the size. Hence, a possible overestimation of salt bridge stability in A99 is not expected to be a major issue for describing the conformational ensemble of the short IDPs studied in this work. This emphasizes the importance of benchmarking against IDPs of different length and sequence when developing and evaluating force fields. While a reduction of the strength of salt bridges appears to be a crucial step in improving the performance of C36, it appears less critical in A99. However, note that this statement is based only on the global conformational properties and that it might be different for studies of dynamics. Based on observations that many force fields have a tendency to overstabilize salt bridges, which seems to be related to side-chain partial charges [22,34–36], we suggest that readjusting the side-chains' partial charges, especially of the phosphorylated residues, is a way of improving the force fields.

Another area which has not been touched upon in this work is the influence of charge regulation and pH. The simulations have been performed with fixed charges in a state corresponding to physiological pH, where the phosphorylated residues have have a charge of $-2e$. Since the pKa of the phosphorylated residues is around six [37], in reality it can fluctuate between $-1e$ and $-2e$. Recent studies have suggested the importance of the protonation state of phosphorylated residues for molecular interactions [38], hence influencing salt bridge formation and the conformational ensemble. Therefore, this is suggested to be included in future investigations.

Considering the secondary structure, the only general difference between the force fields was a higher content of bends in C36. In Tau2, it was focused on regions between salt-bridging-forming partners, suggesting that highly stable salt bridges can enforce bends depending on the separation between the salt-bridging residues. For Tau2, it was suggested that both force fields underestimated the helical propensity, and in Stath, a lack of helix propensity in the N-terminal regions was concerning for C36. However, to properly assess the performance of force fields regarding the secondary structure, detailed experimental references are important. Hence, we see that NMR experiments of phosphorylated IDPs recording coupling constants, NOEs, and chemical shifts, which capture the effects of both the secondary structure and salt bridges, are an essential part of improving force fields. Since atomistic simulations can be used to carefully detect the secondary structure and salt bridges and their dynamics, it is an important tool in understanding the mechanism behind the regulation of IDP function by phosphorylation, provided that sufficient accuracy of the force fields is achieved.

4. Materials and Methods

Fraction of charged residues and $\varkappa$, a parameter describing how segregated the charged residues are in the sequence [31] were calculated in CIDER [39], by equalizing the phosphorylated residues to other negatively charged residues. The value of $\varkappa$ is normalized against the most segregated sequence for that sequence composition, therefore adopting a value in the range 0–1, where 1 corresponds to the most segregated sequence possible.

The simulations listed in Supplementary Table S3 were performed in GROMACS 2018.4 [40–44], using two different force fields: Amber ff99SB-ILDN [23] with the TIP4P-D [24] water model and parameters for the phosphorylated residues from Homeyer et al. [25] and Steinbrecher et al. [26], and CHARMM36m [27] with the CHARMM-modified TIP3P water model [28]. Initial configurations of the peptides were constructed from the sequence as linear chains using Avogadro 1.2.0 [45], optimizing the structure with the auto-optimization tool. Each peptide was solvated in a rhombic dodecahedron box, having a minimum distance between the peptide and the box edges of 1 nm. Sodium ions were added to neutralize the system, and two systems were also simulated with sodium and chloride ions in a concentration corresponding to 150 mM. Periodic boundary conditions were employed in all directions. The equations of motion were integrated using the Verlet leapfrog algorithm [46] with a time step of 2 fs. Nonbonded interactions were treated with a Verlet list cutoff scheme. The short-range interactions were calculated using neighbor lists with cutoff 1 nm or 1.2 nm, for the Amber and CHARMM force fields, respectively. For the CHARMM force field, the Lennard–Jones interactions were switched off smoothly (force-switch) between 1 nm and 1.2 nm. Long-range dispersion corrections were applied to energy and pressure in the case of the Amber force field. Long-range electrostatic interactions were treated by particle mesh Ewald [47] with a cubic interpolation and a 1.6 Å grid spacing. The LINCS algorithm [48] was used to constrain all bond lengths in the case of Amber and only bonds with hydrogen atoms in the case of CHARMM. The solute and solvent were separately coupled to temperature baths at 298 K using the velocity rescaling thermostat [49] with a 0.1 ps relaxation time. Parrinello–Raman pressure coupling [50] was used to keep the pressure at 1 bar, using a 2 ps relaxation time and $4.5 \cdot 10^{-5}$ bar^{-1} isothermal compressibility.

Energy minimization was performed by the steepest descent algorithm until the system converged within the available machine precision. Initiation of five replicates per system with different starting seeds was performed separately in two steps using position restraints on the peptide. The first step was 500 ps of NVT simulation (constant number of particles, volume, and temperature) performed to stabilize the temperature, followed by the second step of 1000 ps of NPT simulation (constant number of particles, pressure, and temperature) to stabilize the pressure. Production runs of the five replicates per system were performed in the NPT ensemble, for at least 1 µs per replicate. The total simulation time per system is stated in Supplementary Table S3. Energies and coordinates were saved every 10 ps. Supplementary Tables S4 and S5 compile a few differences applied to the salt simulations to reduce the computational time.

Analysis

The convergence and sampling quality were assessed in the following ways. The time evolution of the R_g and the R_{ee} in the simulations were observed for signs of equilibration in the initial stage. Based on this, the first 30 ns were removed from each replicate of bCPP in CHARMM36m and the first 50 ns of each replicate of Tau2 in CHARMM36m before final analysis (see Supplementary Figures S21 and S24). In other systems the equilibration was deemed fast enough to be negligible. The distributions of the R_g and the R_{ee} as well as the energy landscapes were compared between replicates, since similarity indicates sufficient sampling. The autocorrelation function and block average error estimates of the R_g and the R_{ee} in the concatenated simulation were calculated and observed for an estimate of the correlation time and convergence of the error estimates. All this data is presented in the Supplementary Figures S6–S33. Although some systems showed greater dissimilarity between replicates than desired, based on the assessment of the concatenated trajectory, it was deemed sufficiently sampled to allow for a comparison between the force fields.

R$_g$ and R$_{ee}$ were calculated using GROMACS 2018.4 [40–44]. Reported error estimates were calculated using block averaging analysis as implemented in the *gmx analyze* routine in GROMACS. Scattering curves were calculated using CRYSOL Version 2.8.3 [51] with the contrast of the hydration shell being 0.0075 $e/\text{Å}^3$ for Amber ff99SB-ILDN+TIP4P-D and 0.02 $e/\text{Å}^3$ for CHARMM36m, as suggested by [29]. The presented curve is the average over 10,000 equally spaced frames. In Supplementary Figure S1 and Table S2, the effect of different contrasts of the hydration shell is shown for Stath. The quality of fit to the experimental curve is computed as:

$$\chi^2(f,c) = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{\text{ref}}(q_i) - (f I_{\text{obs}}(q_i) + c)}{\sigma_{\text{ref}}(q_i)} \right]^2, \tag{1}$$

where N_q^{-1} is the number of points in the reference curve, I_{ref} and I_{obs} are the reference and observed intensities, respectively, and $\sigma_{\text{ref}}(q_i)$ is the error associated with each data point of the reference curve. The function was minimized using the Nelder–Mead method [52], as implemented in Scipy [53], using linear interpolation to produce I_{obs} at the same q points as the reference [29]. AUTORG in the ATSAS program [54] was used to determine the R$_g$ from Guinier analysis. The secondary structure was determined using the DSSP program Version 2.2.1 [55] with an extension to detect the polyproline type II structure [56,57]. The MDTraj Python library Version 1.9.3 [58] was used to calculate contact probability and analyze salt bridges. Contact between two residues was defined as when the shortest distance between two atoms < 0.4 nm. Since salt bridges are formed as a result of hydrogen bonding and electrostatic interactions, they were assessed by analyzing the presence of hydrogen bonds based on the criterion in [59], as implemented in MDTraj. Energy landscapes were calculated following the Campos and Baptista approach [60], with the differences described by Henriques et al. [61]. In short, principal component analysis was applied to the Cartesian coordinates of the backbone atoms of the protein, obtained after translational and rotational least squares fitting on the central structure of the simulation. The conditional free energy was calculated from the probability density function in the representation space constructed by the first two principal components, obtained by Gaussian kernel density estimation. The basins and minima were assigned as described by Campos and Baptista [60]. It is worth noting that the first two components were shown to account for 46–60% of the variance, hence not providing a complete picture of the conformational classes, but at least an overview sufficient for comparison between the force fields. Snapshots from the simulations were produced using VMD 1.9.3 [62–64].

Supplementary Materials: Supplementary material is available.

Funding: Financial support was given by the Royal Physiographic Society in Lund and the Bertil Andersson foundation.

Acknowledgments: The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Center for Scientific and Technical Computing at Lund University (LUNARC).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

A99	Amber ff99SB-ILDN with TIP4P-D water
C36	CHARMM36m with CHARMM-modified TIP3P water
FRET	Fluorescence resonance energy transfer
IDP	Intrinsically disordered protein
NMR	Nuclear magnetic resonance
PPII	polyproline type II
R_g	Radius of gyration
R_{ee}	End-to-end distance
SAXS	Small-angle X-ray scattering

References

1. Dunker, A.; Lawson, J.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001**, *19*, 26–59. doi:10.1016/S1093-3263(00)00138-8.
2. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533. doi:10.1016/S0968-0004(02)02169-2.
3. Fisher, C.K.; Stultz, C.M. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426–431. doi:10.1016/j.sbi.2011.04.001.
4. Johnson, L.N.; Lewis, R.J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **2001**, *101*, 2209–2242. doi:10.1021/cr000225s.
5. Gong, C.X.; Iqbal, K. Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease. *Curr. Med. Chem.* **2008**, *15*, 2321–2328. doi:10.2174/092986708785909111.
6. Raj, P.A.; Johnsson, M.; Levine, M.J.; Nancollas, G.H. Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization. *J. Biol. Chem.* **1992**, *267*, 5968–5976.
7. Makrodimitris, K.; Masica, D.L.; Kim, E.T.; Gray, J.J. Structure Prediction of Protein–Solid Surface Interactions Reveals a Molecular Recognition Motif of Statherin for Hydroxyapatite. *J. Am. Chem. Soc.* **2007**, *129*, 13713–13722. doi:10.1021/ja074602v.
8. De Kruif, C.G.; Holt, C. Casein Micelle Structure, Functions and Interactions. In *Advanced Dairy Chemistry—1 Proteins: Part A/Part B*; Fox, P.F., McSweeney, P.L.H., Eds.; Springer: Boston, MA, USA, 2003; pp. 233–276. doi:10.1007/978-1-4419-8602-3_5.
9. Martin, E.W.; Holehouse, A.S.; Grace, C.R.; Hughes, A.; Pappu, R.V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138*, 15323–15335. doi:10.1021/jacs.6b10272.
10. Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J.D. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* **2010**, *18*, 494–506. doi:10.1016/j.str.2010.01.020.
11. Chin, A.; Toptygin, D.; Elam, W.; Schrank, T.; Hilser, V. Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. *Biophys. J.* **2016**, *110*, 362–371. doi:10.1016/j.bpj.2015.12.013.
12. Schwalbe, M.; Kadavath, H.; Biernat, J.; Ozenne, V.; Blackledge, M.; Mandelkow, E.; Zweckstetter, M. Structural Impact of Tau Phosphorylation at Threonine 231. *Structure* **2015**, *23*, 1448–1458. doi:10.1016/j.str.2015.06.002.
13. RFarrell, H.; Qi, P.; Wickham, E.; Unruh, J. Secondary Structural Studies of Bovine Caseins: Structure and Temperature Dependence of β-Casein Phosphopeptide (1–25) as Analyzed by Circular Dichroism, FTIR Spectroscopy, and Analytical Ultracentrifugation. *J. Protein Chem.* **2002**, *21*, 307–321. doi:10.1023/A:1019992900455.
14. Brister, M.A.; Pandey, A.K.; Bielska, A.A.; Zondlo, N.J. OGlcNAcylation and Phosphorylation Have Opposing Structural Effects in tau: Phosphothreonine Induces Particular Conformational Order. *J. Am. Chem. Soc.* **2014**, *136*, 3803–3816. doi:10.1021/ja407156m.
15. Cragnell, C.; Rieloff, E.; Skepö, M. Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions. *J. Mol. Biol.* **2018**, *430*, 2478–2492. doi:10.1016/j.jmb.2018.03.006.
16. Sieradzan, A.K.; Bogunia, M.; Mech, P.; Ganzynkowicz, R.; Gie?do?, A.; Liwo, A.; Makowski, M. Introduction of Phosphorylated Residues into the UNRES Coarse-Grained Model: Toward Modeling of Signaling Processes. *J. Phys. Chem. B* **2019**, *123*, 5721–5729. doi:10.1021/acs.jpcb.9b03799.
17. Sieradzan, A.K.; Korneev, A.; Begun, A.; Kachlishvili, K.; Scheraga, H.A.; Molochkov, A.; Senet, P.; Niemi, A.J.; Maisuradze, G.G. Investigation of Phosphorylation-Induced Folding of an Intrinsically Disordered Protein by Coarse-Grained Molecular Dynamics. *J. Chem. Theory Comput.* **2021**, *17*, 3203–3220. doi:10.1021/acs.jctc.1c00155.
18. Chong, S.H.; Chatterjee, P.; Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Annu. Rev. Phys. Chem.* **2017**, *68*, 117–134. doi:10.1146/annurev-physchem-052516-050843.
19. Huang, J.; MacKerell, A.D. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2018**, *48*, 40–48. doi:10.1016/j.sbi.2017.10.008.
20. Rieloff, E.; Skepö, M. Phosphorylation of a Disordered Peptide–Structural Effects and Force Field Inconsistencies. *J. Chem. Theory Comput.* **2020**, *16*, 1924–1935. doi:10.1021/acs.jctc.9b01190.
21. Jin, F.; Gräter, F. How multisite phosphorylation impacts the conformations of intrinsically disordered proteins. *PLoS Comput. Biol.* **2021**, *17*, e1008939. doi:10.1371/journal.pcbi.1008939.
22. Ahmed, M.C.; Papaleo, E.; Lindorff-Larsen, K. How well do force fields capture the strength of salt bridges in proteins? *PeerJ* **2018**, *6*, e4967. doi:10.7717/peerj.4967.

23. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958. doi:10.1002/prot.22711.
24. Piana, S.; Donchev, A.G.; Robustelli, P.; Shaw, D.E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123. doi:10.1021/jp508971m.
25. Homeyer, N.; Horn, A.H.C.; Lanig, H.; Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281–289. doi:10.1007/s00894-005-0028-4.
26. Steinbrecher, T.; Latzer, J.; Case, D.A. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory Comput.* **2012**, *8*, 4405–4412. doi:10.1021/ct300613v.
27. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.L.; Grubmüller, H.; MacKerell, A.D., Jr. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2016**, *14*, 71–73.
28. MacKerell, A.D.; Bashford, D.; Bellott, M.; Dunbrack, R.L.; Evanseck, J.D.; Field, M.J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616. doi:10.1021/jp973084f.
29. Henriques, J.; Arleth, L.; Lindorff-Larsen, K.; Skepö, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* **2018**, *430*, 2521–2539. doi:10.1016/j.jmb.2018.03.002.
30. Naganagowda, G.A.; Gururaja, T.L.; Levine, M.J. Delineation of Conformational Preferences in Human Salivary Statherin by 1H, 31P NMR and CD Studies: Sequential Assignment and Structure-Function Correlations. *J. Biomol. Struct. Dyn.* **1998**, *16*, 91–107. doi:10.1080/07391102.1998.10508230.
31. Das, R.K.; Pappu, R.V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 13392–13397. doi:10.1073/pnas.1304749110.
32. Robustelli, P.; Piana, S.; Shaw, D.E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4758–E4766. doi:10.1073/pnas.1800690115.
33. Chan-Yao-Chong, M.; Deville, C.; Pinet, L.; van Heijenoort, C.; Durand, D.; Ha-Duong, T. Structural Characterization of N-WASP Domain V Using MD Simulations with NMR and SAXS Data. *Biophys. J.* **2019**, *116*, 1216–1227. doi:10.1016/j.bpj.2019.02.015.
34. Piana, S.; Lindorff-Larsen, K.; Shaw, D.E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100*, L47–L49. doi:10.1016/j.bpj.2011.03.051.
35. Debiec, K.T.; Gronenborn, A.M.; Chong, L.T. Evaluating the Strength of Salt Bridges: A Comparison of Current Biomolecular Force Fields. *J. Phys. Chem. B* **2014**, *118*, 6561–6569. doi:10.1021/jp500958r.
36. Best, R.B. Atomistic Force Fields for Proteins. In *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Springer: New York, NY, USA, 2019; pp. 3–19. doi:10.1007/978-1-4939-9608-7_1.
37. Bienkiewicz, E.A.; Lumb, K.J. Random-coil chemical shifts of phosphorylated amino acids. *J. Biomol. NMR* **1999**, *15*, 203–206. doi:10.1023/A:1008375029746.
38. Kawade, R.; Kuroda, D.; Tsumoto, K. How the protonation state of a phosphorylated amino acid governs molecular recognition: insights from classical molecular dynamics simulations. *FEBS Lett.* **1999**, *594*, 903–912. doi:10.1002/1873-3468.13674.
39. Holehouse, A.S.; Das, R.K.; Ahad, J.N.; Richardson, M.O.G., Pappu, R.V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins *Biophys. J.* **2017**, *112*, 16–21. doi:10.1016/j.bpj.2016.11.3200.
40. Berendsen, H.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56. doi:10.1016/0010-4655(95)00042-E.
41. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447. doi:10.1021/ct700301q.
42. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M.R.; Smith, J.C.; Kasson, P.M.; van der Spoel, D.; et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854. doi:10.1093/bioinformatics/btt055.
43. Páll, S.; Abraham, M.J.; Kutzner, C.; Hess, B.; Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. In *Solving Software Challenges for Exascale*; Markidis, S., Laure, E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 3–27.
44. 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. doi:10.1016/j.softx.2015.06.001.
45. Hanwell, M.D.; Curtis, D.E.; Lonie, D.C.; Vandermeersch, T.; Zurek, E.; Hutchison, G.R. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **2012**, *4*, 17. doi:10.1186/1758-2946-4-17.
46. Hockney, R.W.; Eastwood, J.W.. Computer Simulation Using Particles. McGraw-Hill: New York, 1981.
47. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. doi:10.1063/1.464397.
48. Hess, B.; Bekker, H.; Berendsen, H.J.C.; Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
49. Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101. doi:10.1063/1.2408420.

50. Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190. doi:10.1063/1.328693.
51. Svergun, D.; Barberato, C.; Koch, M.H.J. *CRYSOL*—A Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768–773. doi:10.1107/S0021889895007047.
52. Nelder, J.A.; Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **1965**, *7*, 308–313. doi:10.1093/comjnl/7.4.308.
53. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. doi:10.1038/s41592-019-0686-2.
54. Franke, D.; Petoukhov, M.V.; Konarev, P.V.; Panjkovich, A.; Tuukkanen, A.; Mertens, H.D.T.; Kikhney, A.G.; Hajizadeh, N.R.; Franklin, J.M.; Jeffries, C.M.; et al. *ATSAS 2.8*: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **2017**, *50*, 1212–1225. doi:10.1107/S1600576717007786.
55. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637. doi:10.1002/bip.360221211.
56. Mansiaux, Y.; Joseph, A.P.; Gelly, J.C.; de Brevern, A.G. Assignment of PolyProline II Conformation and Analysis of Sequence—Structure Relationship. *PLoS ONE* **2011**, *6*, 1–15. doi:10.1371/journal.pone.0018401.
57. Chebrek, R.; Leonard, S.; de Brevern, A.G.; Gelly, J.C. PolyprOnline: polyproline helix II and secondary structure assignment database. *Database* **2014**, *2014*. doi:10.1093/database/bau102.
58. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. doi:10.1016/j.bpj.2015.08.015.
59. Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Näslund, L.Å.; Hirsch, T.K.; Ojamäe, L.; Glatzel, P.; et al. The Structure of the First Coordination Shell in Liquid Water. *Science* **2004**, *304*, 995–999. doi:10.1126/science.1096205.
60. Campos, S.R.R.; Baptista, A.M. Conformational Analysis in a Multidimensional Energy Landscape: Study of an Arginylglutamate Repeat. *J. Phys. Chem. B* **2009**, *113*, 15989–16001. doi:10.1021/jp902991u.
61. Henriques, J.; Cragnell, C.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 3420–3431. doi:10.1021/ct501178z.
62. Humphrey, W.; Dalke, A.; Schulten, K. VMD—Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
63. Stone, J. An Efficient Library for Parallel Ray Tracing and Animation. Master's Thesis, Computer Science Department, University of Missouri-Rolla, Rolla, MO, USA, 1998.
64. Frishman, D.; Argos, P. Knowledge-based secondary structure assignment. *Proteins* **1995**, *23*, 566–579.

Supplementary Materials for Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison

E. Rieloff and M. Skepö

This file contains:

- Table S1: Fraction of charged residues and $\varkappa$ of the peptides studied.
- Table S2: Radius of gyration and χ^2 of calculated scattering curves using different contrast of hydration shell ($\delta\rho$) for Stath.
- Table S3: Details of the simulations performed in this work.
- Table S4: Differences in the setup of systems with 150 mM NaCl.
- Table S5: Starting configuration used in simulations of system bCPP 150 mM C36.
- Figure S1: Calculated scattering curves of Stath using different contrast of the hydration shell.
- Figure S2: Content of different secondary structure elements along the sequence for the four peptides.
- Figure S3: Difference in contact probability for bCPP between 0 and 150 mM NaCl.
- Figure S4: Distribution of radius of gyration of end-to-end distance for bCPP with 0 and 150 mM NaCl.
- Figure S5: Calculated scattering curves for bCPP with 0 and 150 mM NaCl.
- Figure S6-S8: Plots for assessing convergence and sampling quality of the simulations of Tau1 in Amber ff99SB-ILDN.
- Figure S9-S11: Plots for assessing convergence and sampling quality of the simulations of Tau2 in Amber ff99SB-ILDN.
- Figure S12-S14: Plots for assessing convergence and sampling quality of the simulations of bCPP in Amber ff99SB-ILDN.
- Figure S15-S17: Plots for assessing convergence and sampling quality of the simulations of Stath in Amber ff99SB-ILDN.
- Figure S18-S20: Plots for assessing convergence and sampling quality of the simulations of Tau1 in CHARMM36m.
- Figure S21-S23: Plots for assessing convergence and sampling quality of the simulations of Tau2 in CHARMM36m.

- Figure S24-S26: Plots for assessing convergence and sampling quality of the simulations of bCPP in CHARMM36m.
- Figure S27-S29: Plots for assessing convergence and sampling quality of the simulations of Stath in CHARMM36m.
- Figure S30-S31: Plots for assessing convergence and sampling quality of the simulations of bCPP with 150 mM NaCl in Amber ff99SB-ILDN.
- Figure S32-S33: Plots for assessing convergence and sampling quality of the simulations of bCPP with 150 mM NaCl in CHARMM36m.

Table S1: Fraction of charged residues (FCR) and level of charge separation described by $\varkappa$ of the peptides studied.

Peptide	FCR	$\varkappa$
Tau1	0.29	0.05
Tau2	0.41	0.12
bCPP	0.52	0.46
Stath	0.23	0.32

Table S2: Radius of gyration and χ^2 of calculated scattering curves using different contrast of hydration shell ($\delta\rho$) for Stath simulated with Amber FF99-SB-ILDN (A99) and CHARMM36m (C36). The R_g is obtained from Guinier analysis using AUTORG in the ATSAS program [1], and the error reported is the estimated error given by AUTORG.

	A99		**C36**	
$\delta\rho(e/\text{Å}^3)$	R_g (Å)	χ^2	R_g (Å)	χ^2
0	17.2 ± 0.6	2.0	13.9 ± 0.4	6.7
0.01	17.5 ± 0.6	1.6	14.2 ± 0.3	5.3
0.02	17.8 ± 0.5	1.4	14.6 ± 0.3	4.4
0.03	18.1 ± 0.5	1.3	14.9 ± 0.3	3.7

Table S3: Details of the simulations performed in this work.

Peptide	Force field	Salt concentration (mM)	Box volume (nm^3)	Number of solvent molecules	Number of sodium ions	Number of chloride ions	Total simulation length (µs)
Tau1	A99	0	263.66	8637	2	0	5
Tau2	A99	0	722.941	23816	3	0	11
bCPP	A99	0	1002.41	32815	13	0	6
Stath	A99	0	942.11	30942	4	0	12
Tau1	C36	0	263.75	8495	2	0	11
Tau2	C36	0	722.93	23519	3	0	8.05
bCPP	C36	0	1002.48	32381	13	0	6.75
Stath	C36	0	950.87	30708	4	0	6
bCPP	A99	150	1002.41	32633	104	91	7
bCPP	C36	150	1002.48	32199	104	91	9.49

*A99 = Amber ff99SB-ILDN with the TIP4P-D water model, C36 = CHARMM36m with the CHARMM-modified TIP3P water model.

Table S4: Differences in the setup of the systems with 150 mM NaCl.

System	Number of replicates	Minimum simulation length of replicate (µs)	Saving frequency (pS)
bCPP 150 mM A99	10	0.7	40
bCPP 150 mM C36	10	0.48	50

Table S5: Starting configuration used in simulations of system bCPP 150 mM C36.

Replicate number	Starting configuration
1	Linear
2	Linear
3	Linear
4	Linear
5	Linear
6	t=58 ns in replicate #1
7	t=58 ns in replicate #2
8	t=58 ns in replicate #3
9	t=58 ns in replicate #4
10	t=58 ns in replicate #5

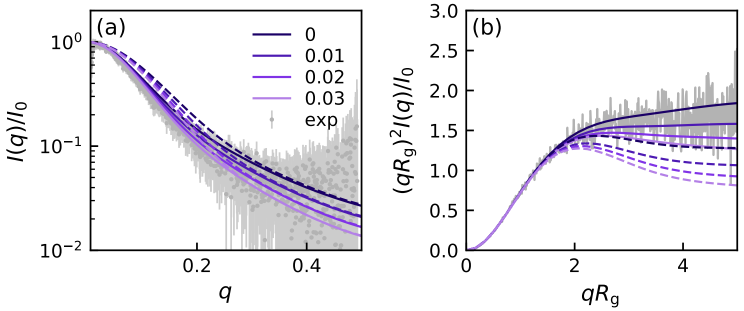


Figure S1: Calculated scattering curves of Stath using different contrast of the hydration shell presented as a semi-log plot (a) and dimensionless Kratky plot (b). Solid lines correspond to Amber ff99SB-ILDN+TIP4P-D and dashed lines to CHARMM36m. The experimental curve is the form factor of Statherin first presented in reference [2].

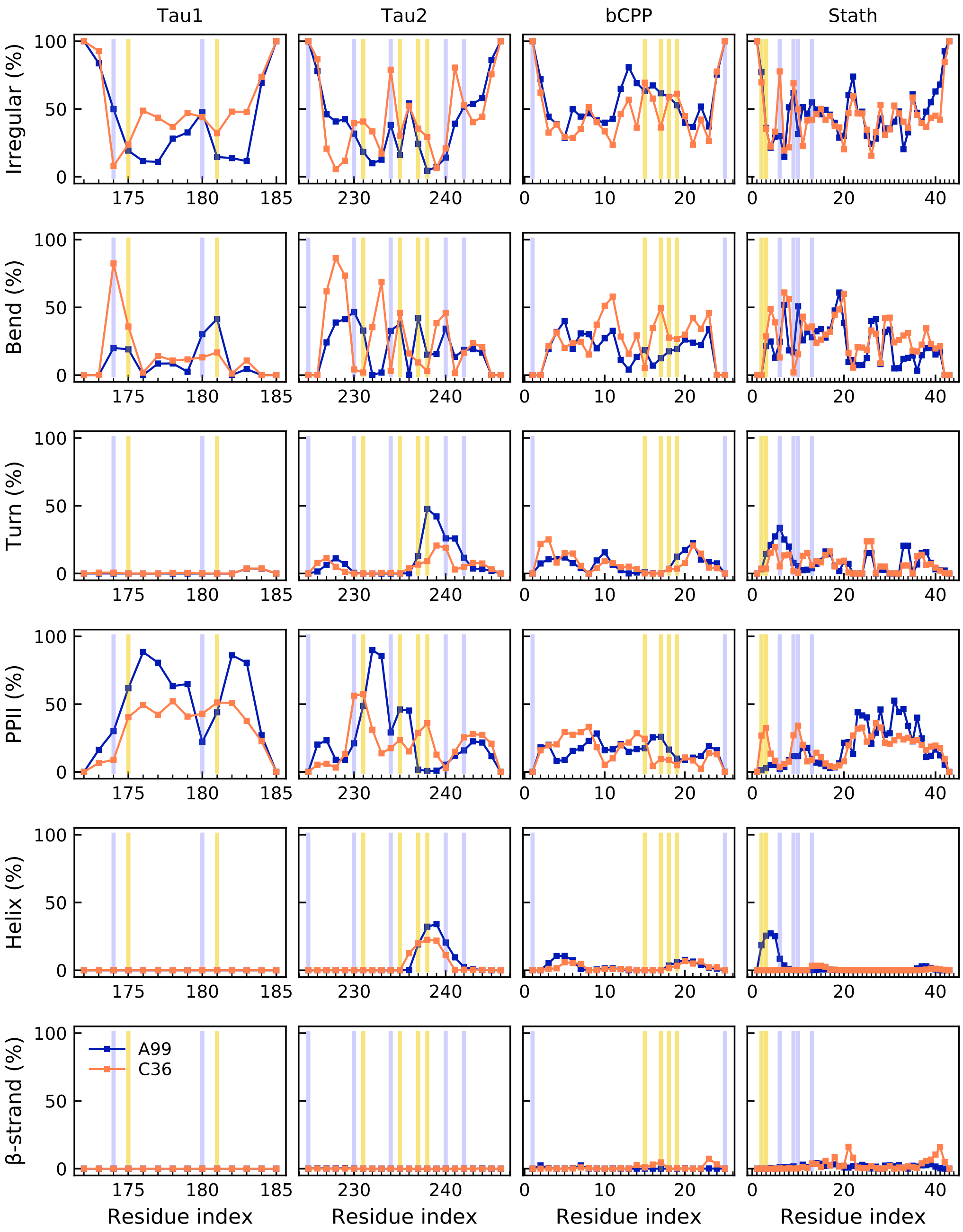


Figure S2: Content of different secondary structure elements along the sequence for the four peptides. The legend in the lower left panel applies to all panels. The position of phosphorylated residues are highlighted in yellow, and the position of positively charged residues in blue. Helix contains both α-, 3_{10}- and π-helix, and β-strand contain also includes β-bridge.

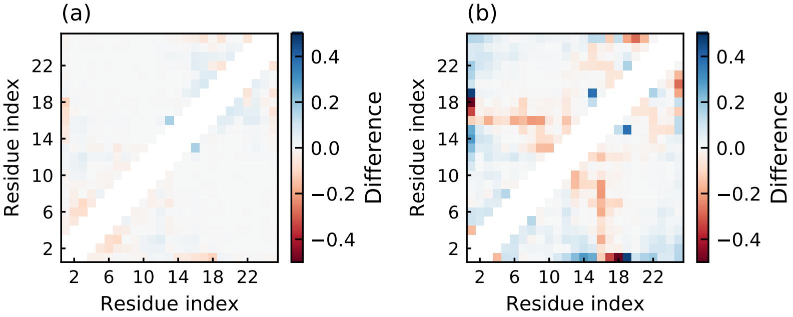


Figure S3: Difference in contact probability between 0 and 150 mM salt for bCPP simulated with Amber ff99SB-ILDN (a) and CHARMM36m (b).

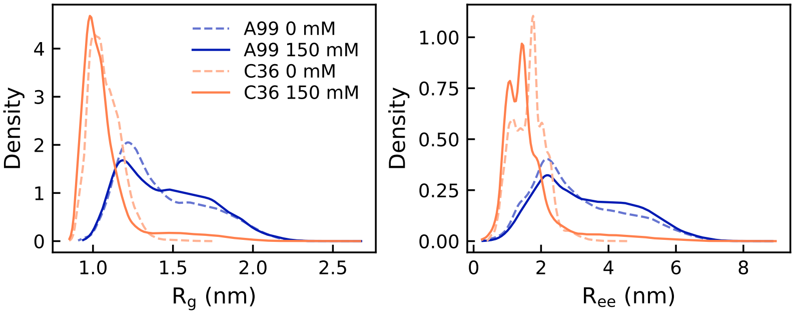


Figure S4: Distribution of radius of gyration (a) and end-to-end distance (b) of bCPP simulated with Amber ff99SB-ILDN (A99) and CHARMM36m (C36) in the presence of 0 or 150 mM NaCl. The legend applies to both panels.

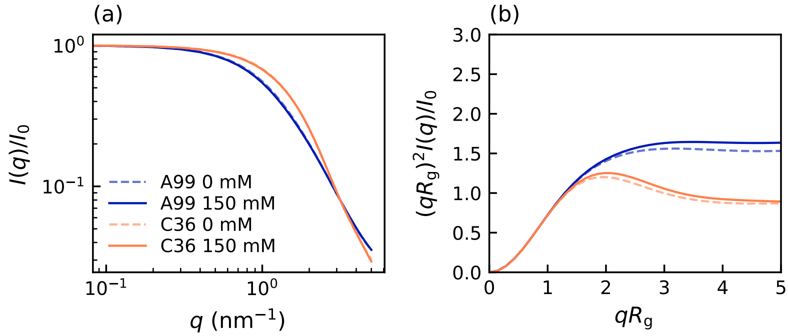


Figure S5: Calculated form factor (a) and dimensionless Kratky representation (b) of bCPP simulated with Amber ff99SB-ILDN (A99) and CHARMM36m (C36) in the presence of 0 or 150 mM NaCl. The legend applies to both panels.

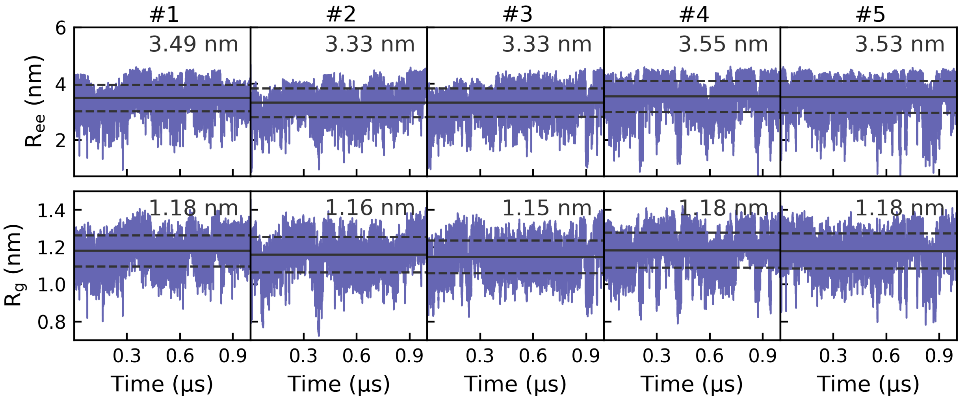


Figure S6: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Tau1 in Amber ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

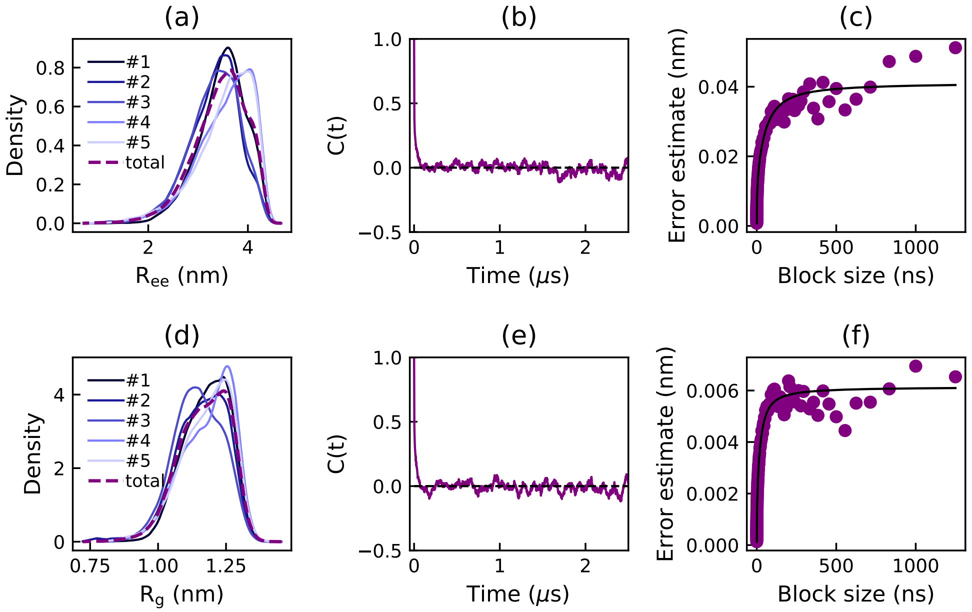


Figure S7: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of Tau1 in Amber ff99SB-ILDN, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

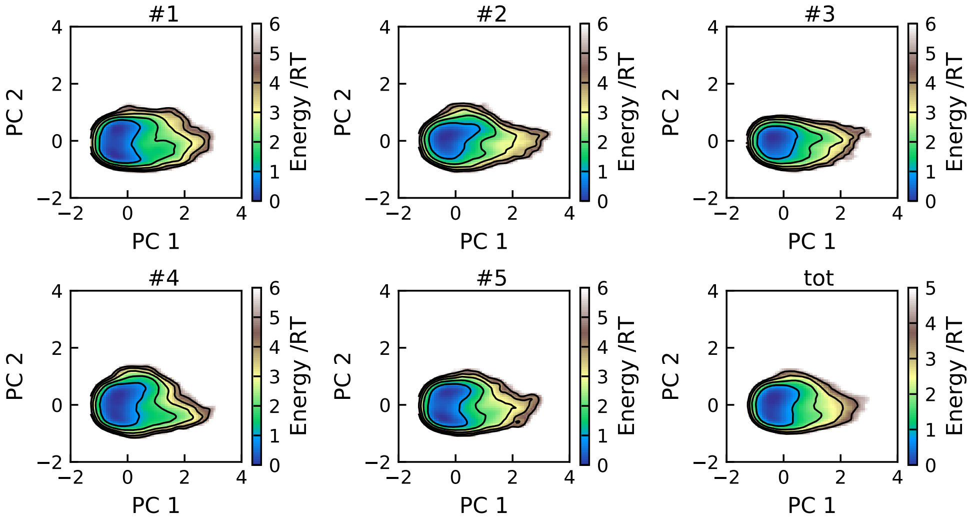


Figure S8: Energy landscapes for the five replicates and the concatenated trajectory of Tau1 in Amber ff99SB-ILDN, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

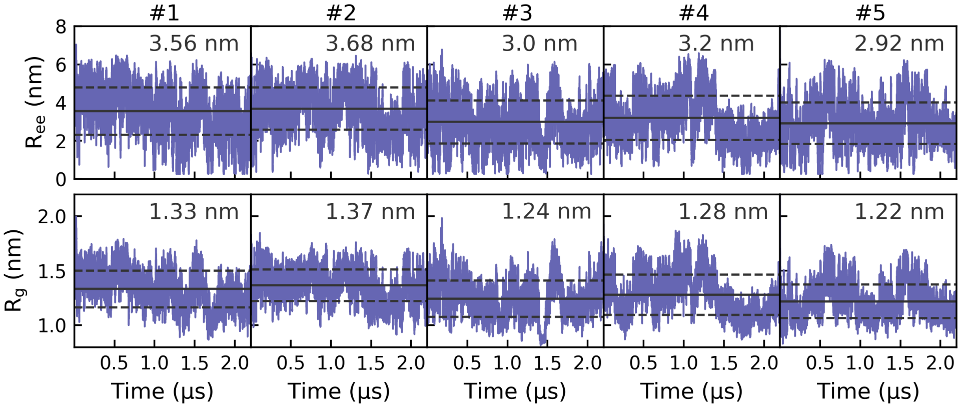


Figure S9: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Tau2 in Amber ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

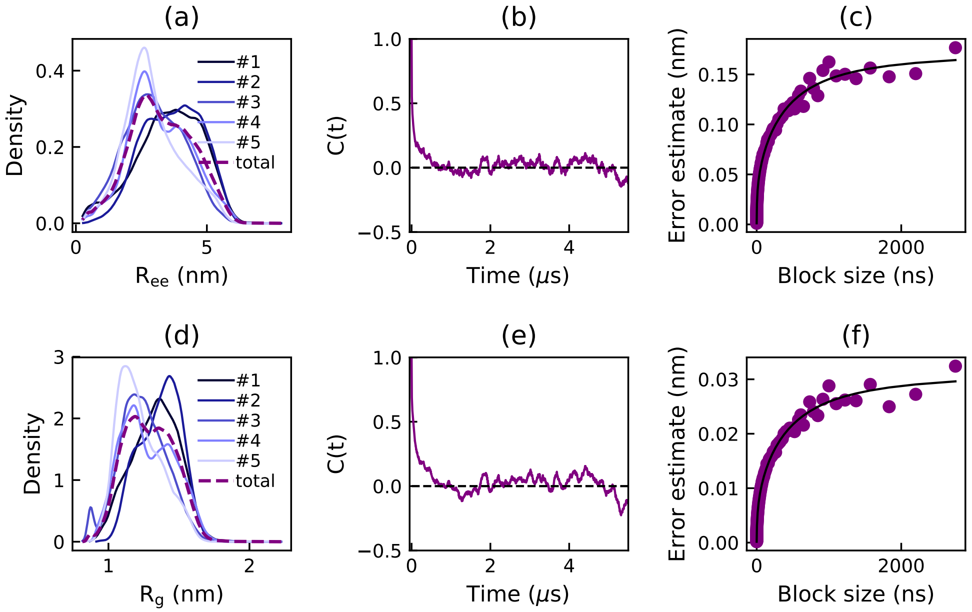


Figure S10: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of Tau2 in Amber ff99SB-ILDN, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

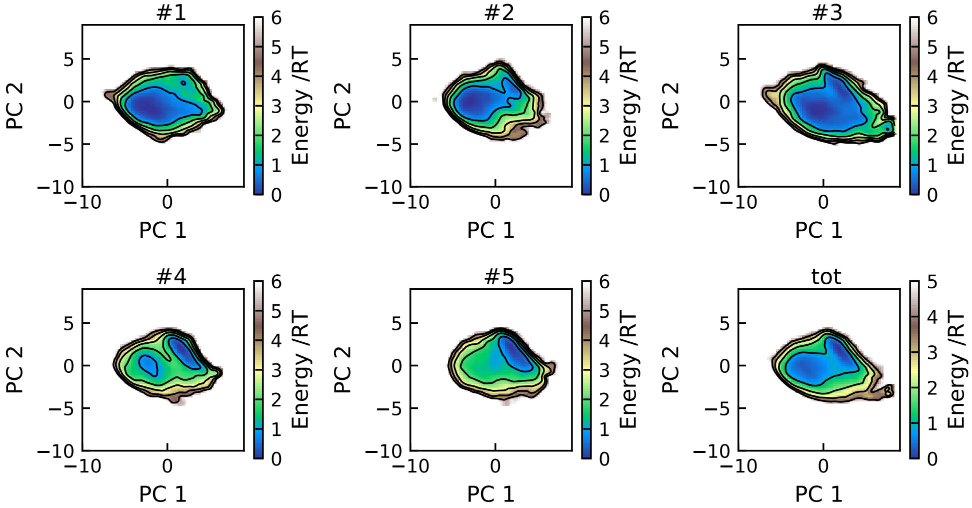


Figure S11: Energy landscapes for the five replicates and the concatenated trajectory of Tau2 in Amber ff99SB-ILDN, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

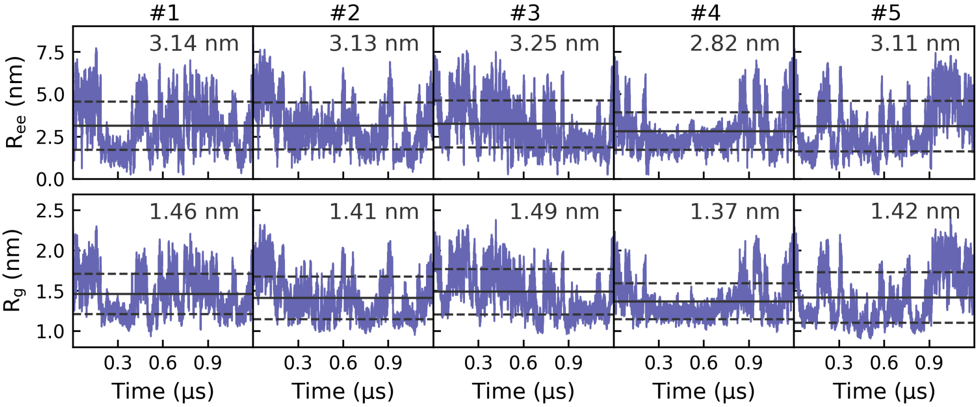


Figure S12: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of bCPP in Amber ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

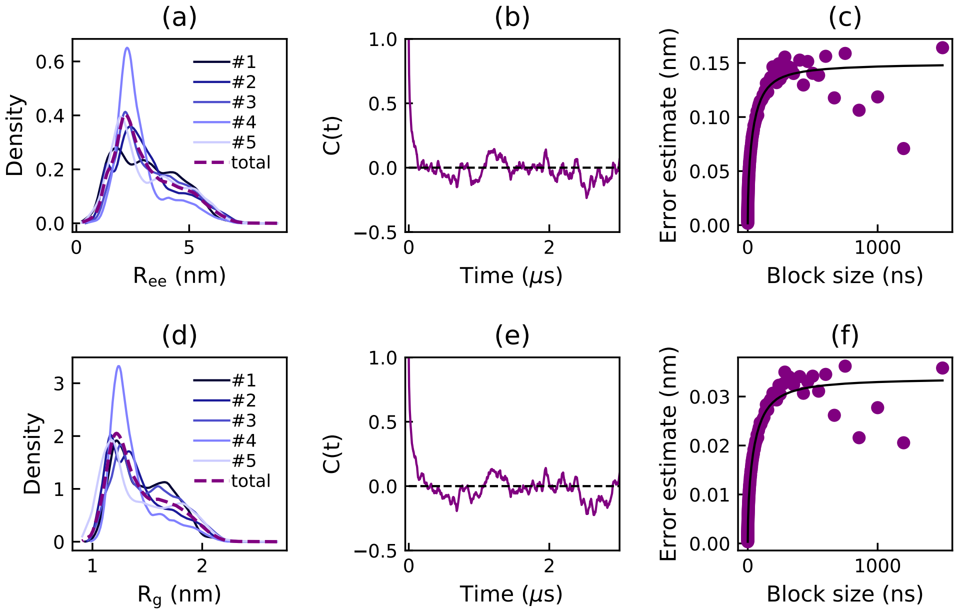


Figure S13: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of bCPP in Amber ff99SB-ILDN, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

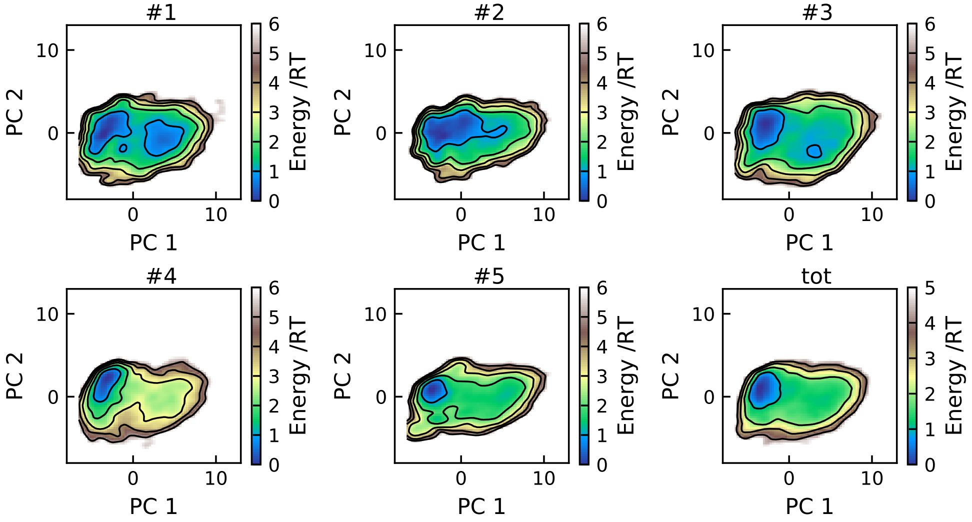


Figure S14: Energy landscapes for the five replicates and the concatenated trajectory of bCPP in Amber ff99SB-ILDN, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

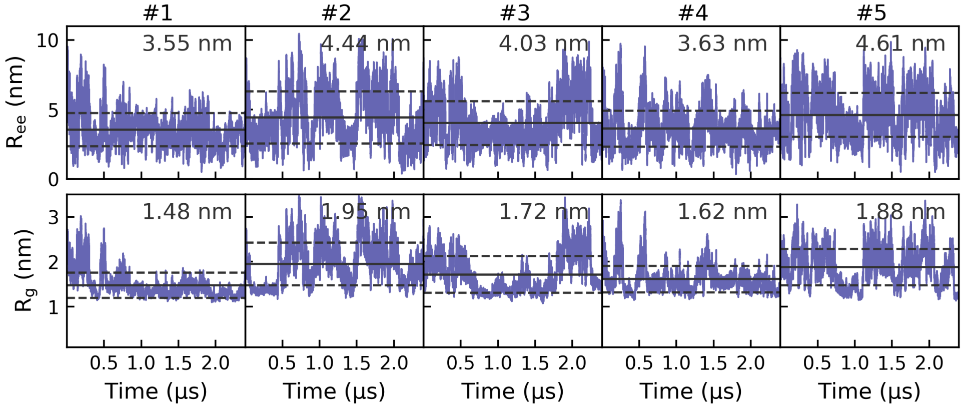


Figure S15: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Stath in Amber ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

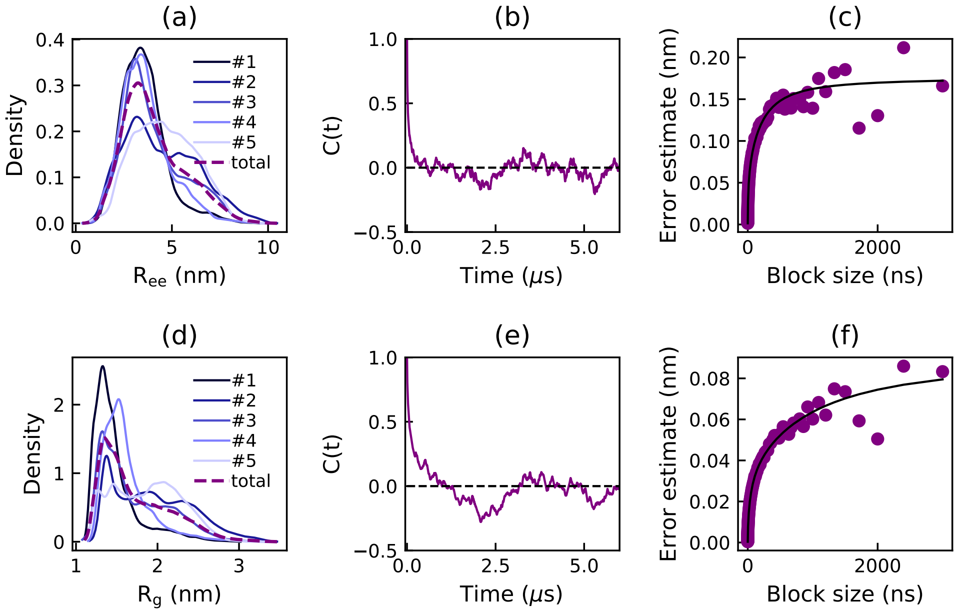


Figure S16: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of Stath in Amber ff99SB-ILDN, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

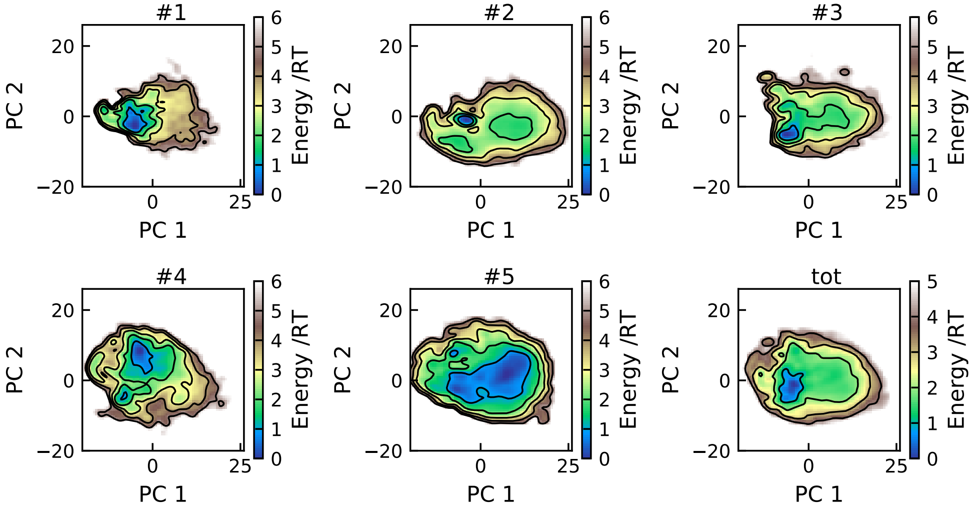


Figure S17: Energy landscapes for the five replicates and the concatenated trajectory of Stath in Amber ff99SB-ILDN, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

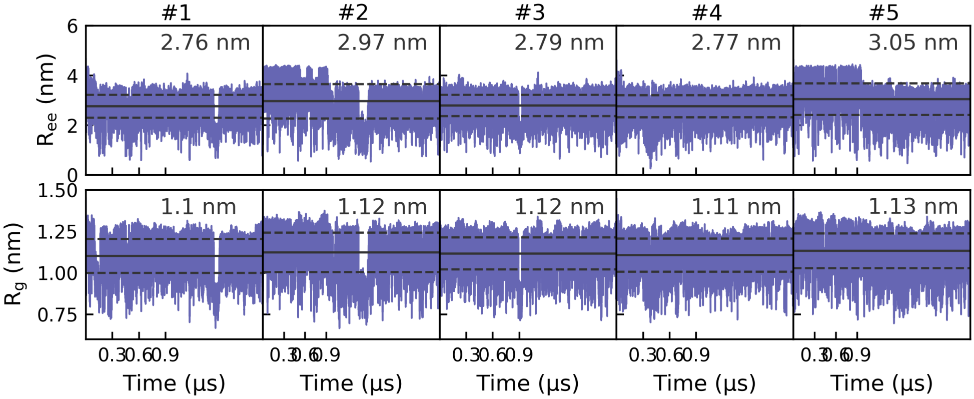


Figure S18: Time evolution of the end-to-end distance ($\mathrm{R_{ee}}$) and the radius of gyration ($\mathrm{R_g}$) for the five replicates in the simulation of Tau1 in CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

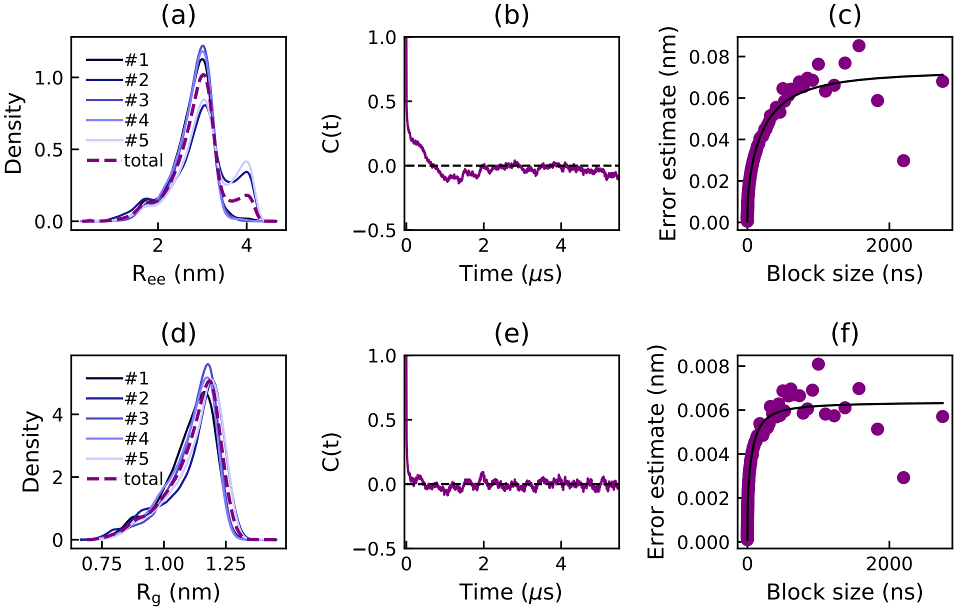


Figure S19: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of Tau1 in CHARMM36m, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

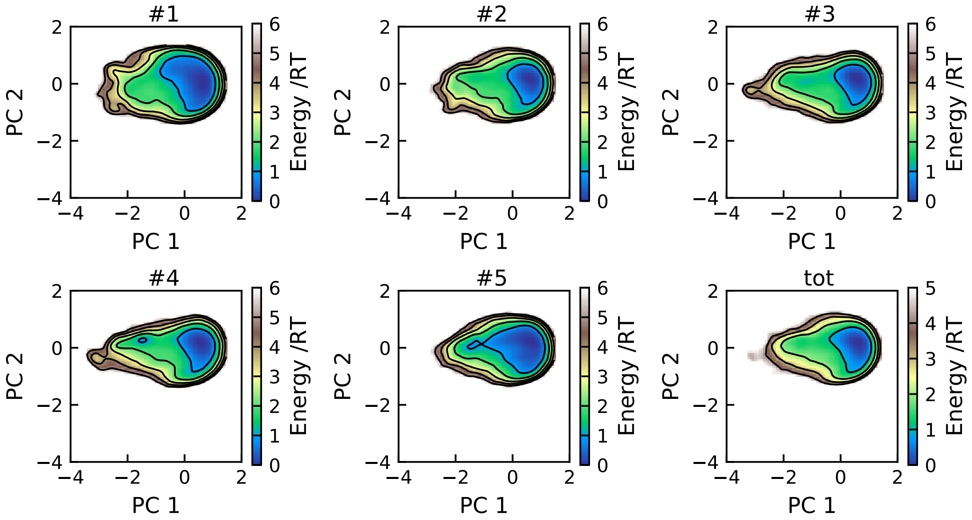


Figure S20: Energy landscapes for the five replicates and the concatenated trajectory of Tau1 in CHARMM36m, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

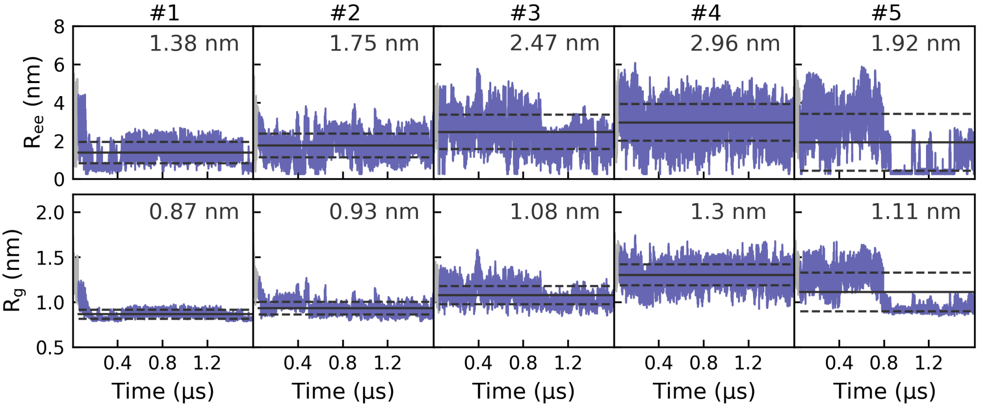


Figure S21: Time evolution of the end-to-end distance ($\mathrm{R_{ee}}$) and the radius of gyration ($\mathrm{R_g}$) for the five replicates in the simulation of Tau2 in CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation. The region removed before final analysis is plotted in gray.

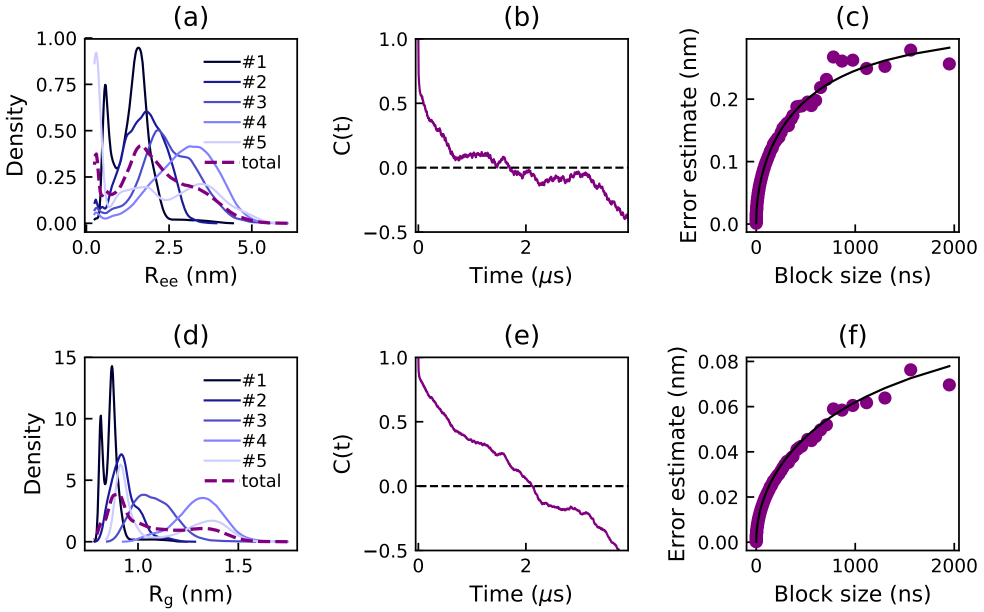


Figure S22: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of Tau2 in CHARMM36m, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

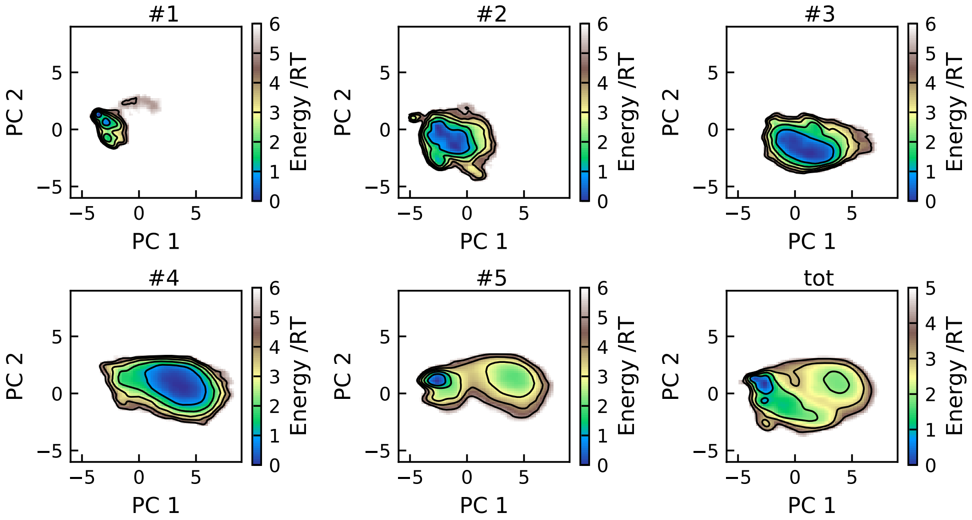


Figure S23: Energy landscapes for the five replicates and the concatenated trajectory of Tau2 in CHARMM36m, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

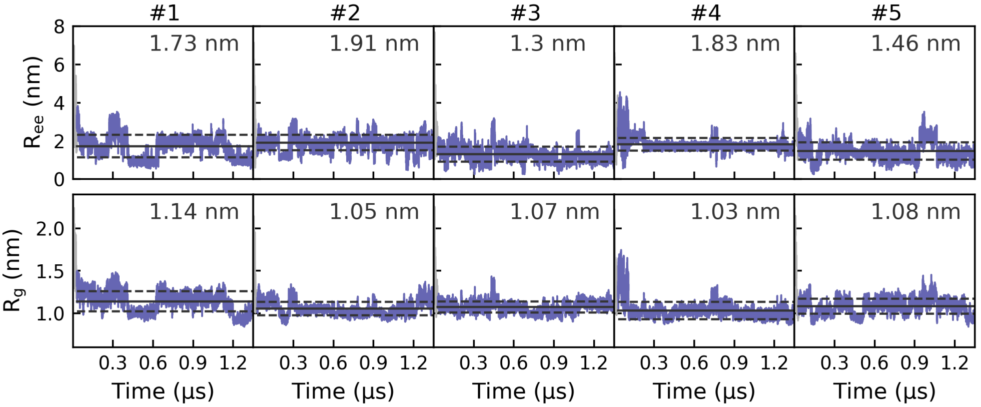


Figure S24: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of bCPP in CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation. The region removed before final analysis is plotted in gray.

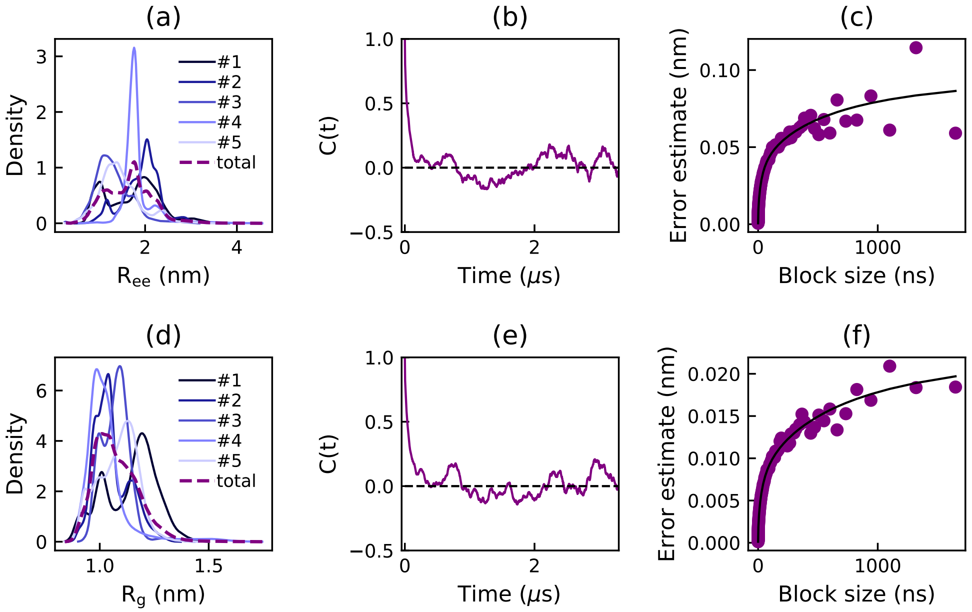


Figure S25: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of bCPP in CHARMM36m, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

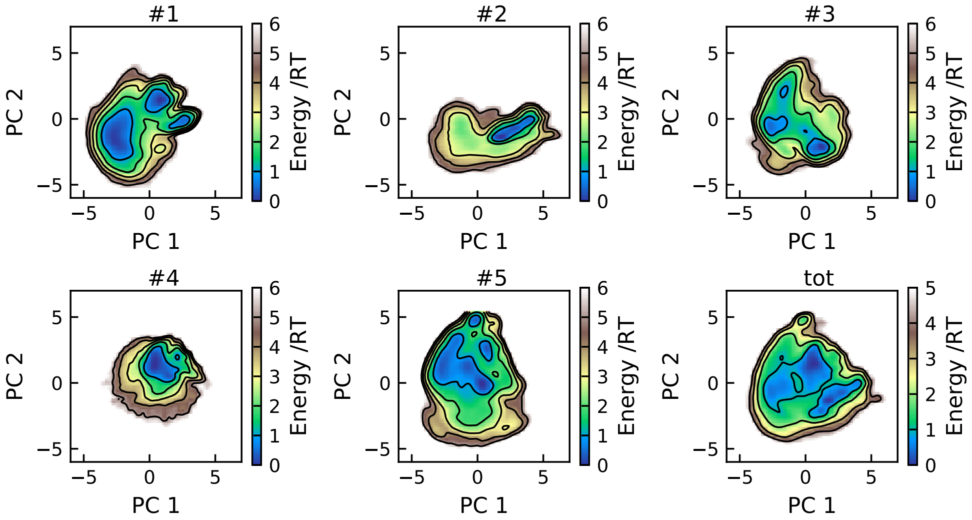


Figure S26: Energy landscapes for the five replicates and the concatenated trajectory of bCPP in CHARMM36m, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

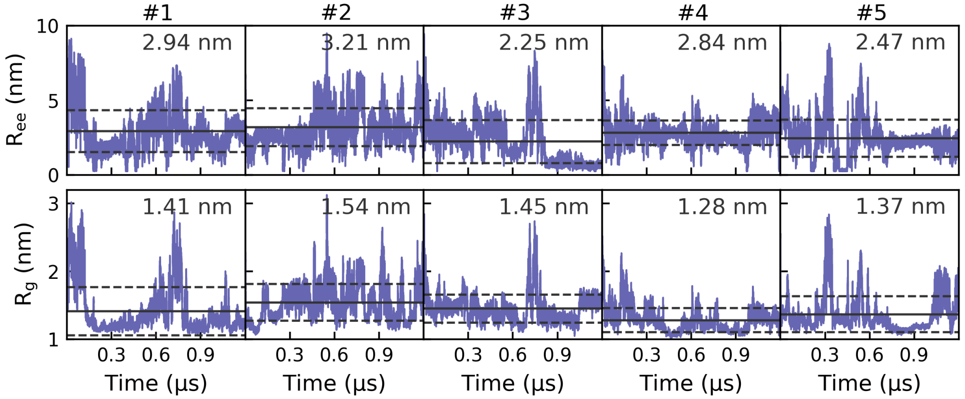


Figure S27: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Stath in CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

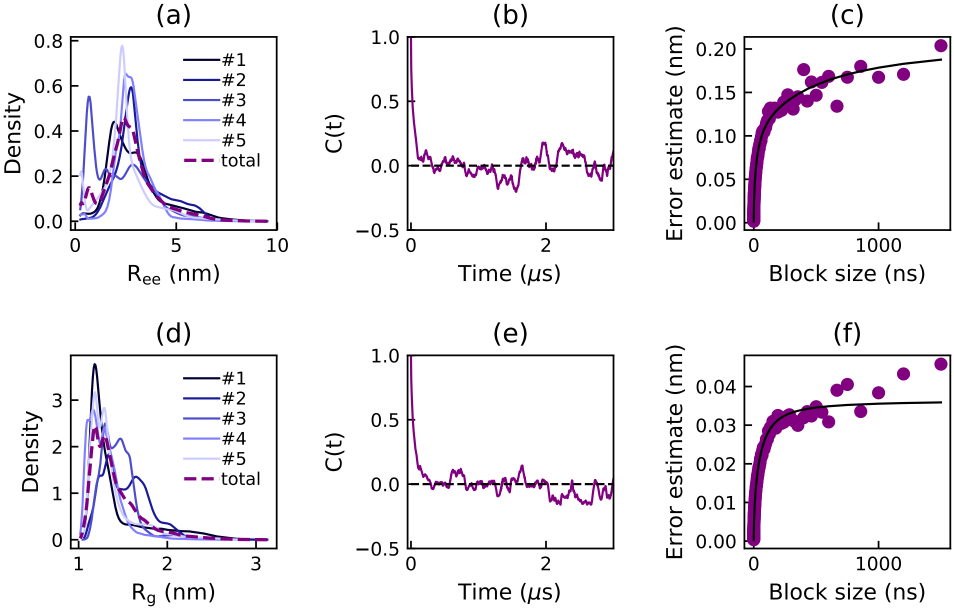


Figure S28: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the five replicates and the concatenated simulation of Stath in CHARMM36m, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

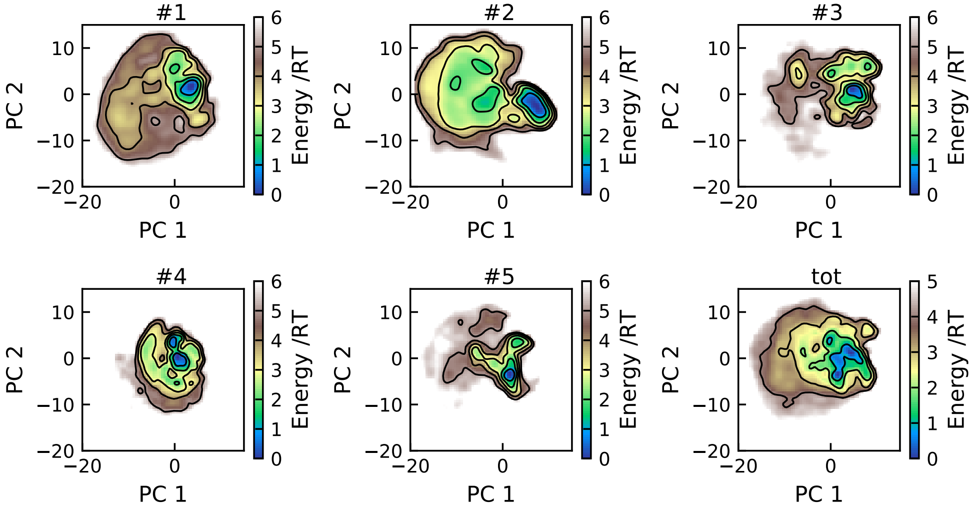


Figure S29: Energy landscapes for the five replicates and the concatenated trajectory of Stath in CHARMM36m, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

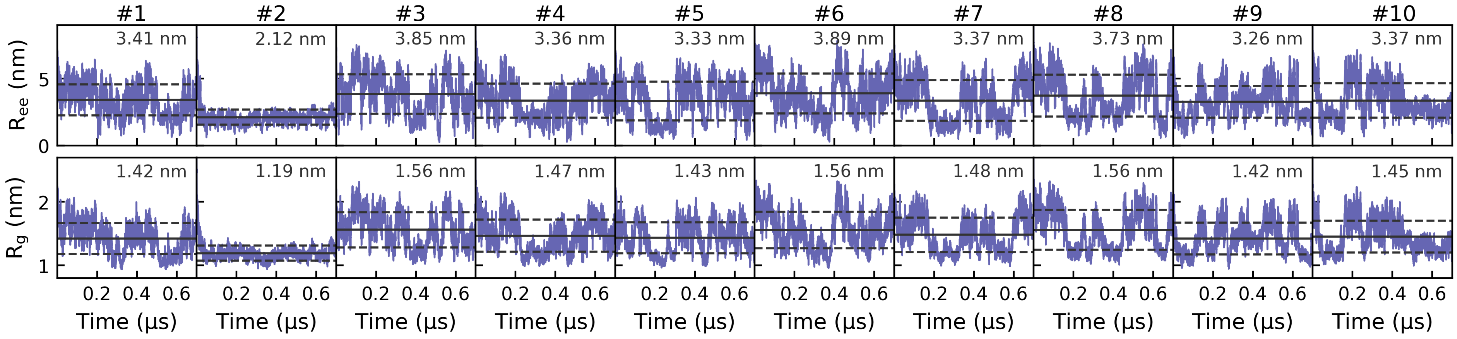


Figure S30: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the ten replicates in the simulation of bCPP with 150 mM NaCl in Amber ff99SB-ILDN. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

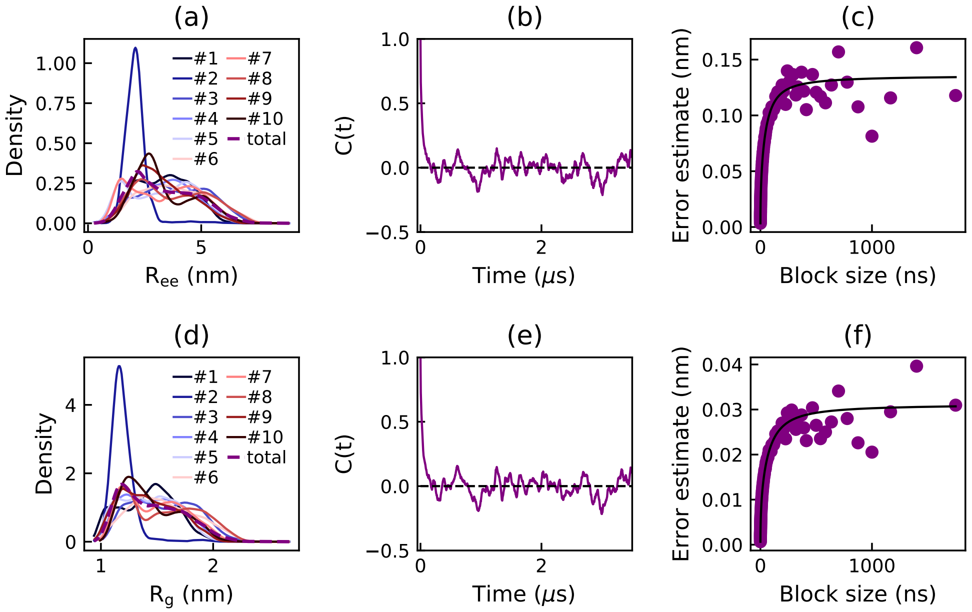


Figure S31: Density estimates of the end-to-end distance (R$_{ee}$) (a) and the radius of gyration (R$_g$) (d) for the ten replicates and the concatenated simulation of bCPP with 150 mM NaCl in Amber ff99SB-ILDN, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

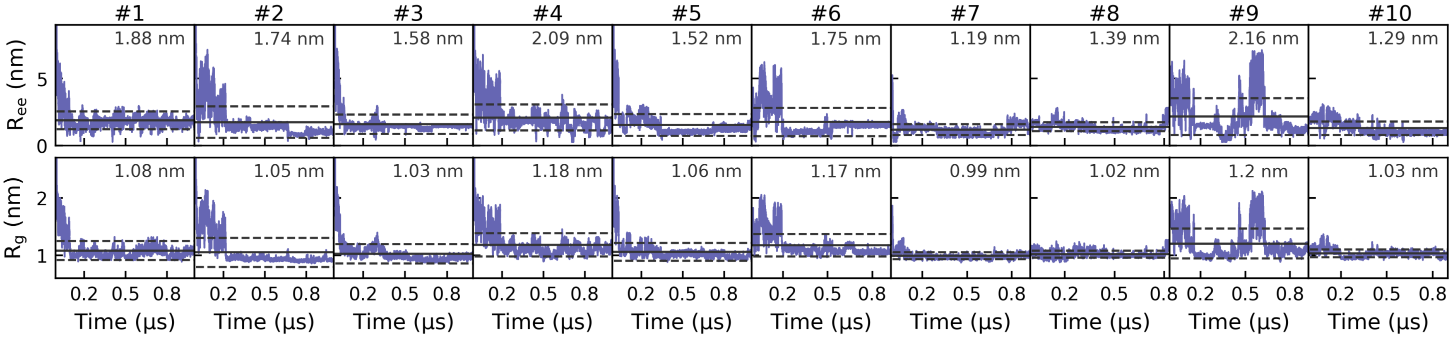


Figure S32: Time evolution of the end-to-end distance (R$_{ee}$) and the radius of gyration (R$_g$) for the ten replicates in the simulation of bCPP with 150 mM NaCl in CHARMM36m. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

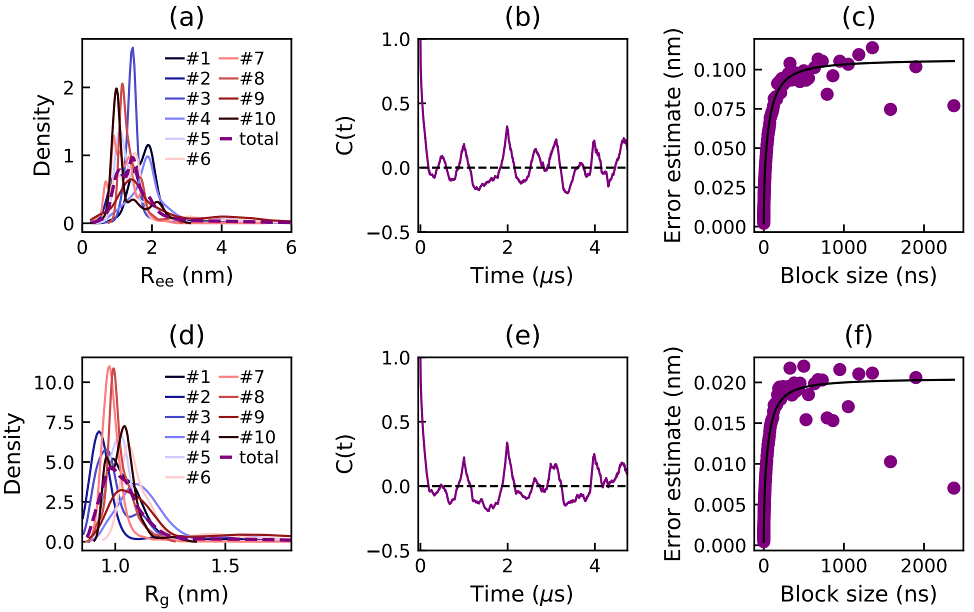


Figure S33: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the ten replicates and the concatenated simulation of bCPP with 150 mM NaCl in CHARMM36m, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

References

[1] D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, and D. I. Svergun. *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.*, 50(4):1212–1225, Aug 2017.

[2] Carolina Cragnell, Ellen Rieloff, and Marie Skepö. Utilizing coarse-grained modeling and monte carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. *Journal of Molecular Biology*, 430(16):2478–2492, 2018.

Paper v

Article

The effect of multisite phosphorylation on the conformational properties of intrinsically disordered proteins

Ellen Rieloff [1] and Marie Skepö [1,2],*

[1] Division of Theoretical Chemistry, Lund University, POB 124, SE-221 00 Lund, Sweden
[2] LINXS – Lund Institute of Advanced Neutron and X-ray Science, Scheelevägen 19, SE-223 70 Lund, Sweden
* Correspondence: marie.skepo@teokem.lu.se

Abstract: Intrinsically disordered proteins are involved in many biological processes such as signaling, regulation, and recognition. A common strategy to regulate their function is through phosphorylation, as it can induce changes in conformation, dynamics, and interactions with binding partners. Although phosphorylated intrinsically disordered proteins have received increased attention in recent years, a full understanding of the conformational and structural implications of phosphorylation has not yet been achieved. Here we have performed all-atom molecular dynamics simulations of five disordered peptides originated from tau, statherin, and β-casein, in both phosphorylated and non-phosphorylated state, to compare changes in global dimensions and structural elements. The changes are in qualitative agreement with experimental data, and we observe that the net charge is not enough to predict the impact of phosphorylation on the global dimensions. Instead, the distribution of phosphorylated and positively charged residues throughout the sequence has great impact due to the formation of salt bridges. In statherin, a preference for arginine–phosphoserine interaction over arginine–tyrosine accounts for a global expansion, despite a local contraction of the phosphorylated region, which implies that also non-charged residues can influence the effect of phosphorylation.

Keywords: intrinsically disordered proteins, phosphorylation, force fields

Citation: Rieloff, E.; Skepö, M. The effect of multisite phosphorylation on the conformational properties of intrinsically disordered proteins. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Intrinsically disordered proteins (IDPs) lack tertiary structure under physiological conditions [1,2], such that they adopt a range of different interchanging conformations rather than a single structure. This is reflected in their rather flat free energy landscapes [3], making them sensitive to environmental changes and post-translational modifications (PTMs), which helps to regulate function. Many IDPs also demonstrate the ability to bind to several targets, and adopt different folds depending on the target. These characteristics of IDPs are advantageous in signaling, regulation, and recognition processes, where IDPs are abundantly involved [4,5].

Phosphorylation is a reversible type of PTM, especially prevalent among intrinsically disordered regions and proteins [6–8]. The addition of a bulky phosphoryl group to residues such as serine or threonine adds extra negative charge and enables formation of hydrogen bonds and salt bridges [9], which can induce drastic changes in the conformational ensemble and the dynamics of the IDP. In a simplistic view, assuming electrostatics to be the major determinant of IDP conformation, a net positively charged IDP is expected to contract upon phosphorylation, while a negatively charged or neutral IDP will expand. In a recent atomistic simulation study by Jin and Gräter this prediction was shown to hold true for multisite phosphorylation of the four peptides studied [10]. Generally, net charge and hydropathy provide good indications of the level of compaction of a protein only in some cases, while many require an additional inspection of the fraction of charged residues and charge pattern, due to their polyampholytic nature [11,12].

In recent years, phosphorylated IDPs have received increased attention [10,13–23]. Changes in global conformation, secondary structure, and local arrangements upon phosphorylation of disordered proteins and regions have been studied experimentally by techniques such as small angle X-ray scattering (SAXS), fluorescence resonance energy transfer, circular dichroism (CD) spectroscopy, and nuclear magnetic resonance (NMR) [13–15,20,24–26]. Due to the vast conformational ensembles possessed by IDPs, a combination of different techniques is required and often well complemented by atomistic simulations, which through detailed information can provide further insight. After many years of important adjustments, such as refinement of backbone dihedral angles and balancing the water–protein and protein–protein interactions, there are several force field and water model combinations that can be applied to IDPs [27,28]. Less attention has been given to charge–charge interactions, although it has been determined that many standard force fields have a tendency to overestimate salt bridges [29,30]. More recently, it has been shown that for phosphorylated peptides this can cause serious discrepancies between simulations and experiments [10,20,31]. However, in our most recent work, Amber ff99SB-ILDN in combination with the TIP4P-D water model have showed promising results in describing the conformational ensemble of short disordered peptides [20,31].

Here we have used all-atom molecular dynamics simulations with the Amber ff99SB-ILDN force field to study the conformational and structural effects upon phosphorylation of five disordered peptides, to gain better insight into the controlling factors. By experimental comparison we also detect limitations of the force field. Two of the peptides are fragments from the neuroprotein tau, involved in stabilizing neuronal microtubules [32]. Phosphorylation of tau regulate its function, and hyperphosphorylation has been implicated to cause pathological effects by involvement in amyloid fibril formation in Alzheimer's disease [33,34]. Another two of the peptides are the saliva protein statherin and its fifteen residue long N-terminal fragment, SN15. Statherin maintains a supersaturated environment of calcium phosphate in the saliva, by preventing spontaneous precipitation and crystal growth [35–37]. This functionality is closely associated with the N-terminal fragment containing the phosphorylated residues [37]. The last peptide is the 25 residue long N-terminal fragment of β-casein, which naturally contains four phosphorylated serines that sequester calcium and promotes the formation of calcium phosphate nanoclusters [38–40].

We observe that for these peptides, ranging in length from 11 to 43 residues, that net charge is not enough to predict the change in global dimensions upon phosphorylation at two to four sites. Instead, salt bridge formation has great impact, depending on the distribution of phosphorylated and positively charged residues throughout the sequence. Further, in statherin, a preference for arginine–phosphoserine interactions over arginine–tyrosine interactions explains the phosphorylation induced changes.

Results and Discussion

Net charge is not enough to explain phosphorylation induced changes

Atomistic simulations of five different disordered peptides in both non-phosphorylated and phosphorylated state, shown in Table 1, have been performed at physiological pH. The peptides were chosen based on the availability of experimental data and their size, considering computational expense.

Table 1. Full name and sequence of the peptides included in this study. Positively charged residues are marked in blue, negatively charged in red, and phosphorylation sites are highlighted with yellow. The number of residues (N_{res}), net charge of the non-phosphorylated variant (Z_{no}) and the phosphorylated variant (Z_{ph}) are also shown.

Name	Peptide	Sequence	N_{res}	Z_{no}	Z_{ph}
Tau1	$Tau_{173–183}$	AKTPPAPKTPP	11	+2	-2
SN15	$Statherin_{1–15}$	DSSEEKFLRRIGRFG	15	+1	-3
Tau2	$Tau_{225–246}$	KVAVVRTPPKSPSSAKSRLQTA	22	+5	-3
bCPP	β-$casein_{1–25}$	RELEELNVPGEIVESLSSSEESITR	25	-5	-13
Stath	Statherin	DSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF	43	0	-4

SN15, Tau2, and bCPP all contract upon phosphorylation, as shown from the peak shift towards lower values of the distributions of radius of gyration (R_g) and end-to-end distance (R_{ee}) in Figure 1, as well as the average values of R_g and R_{ee} presented in Table 2. For SN15 and Tau2 the width of the distribution also decreases, while bCPP keeps the same range, only the shape of the distribution changes. Stath and Tau1 both expand, shown from a peak shift towards larger values in the distributions. For Tau1 the expansion is more clear observing the R_g distribution than the R_{ee} distribution, which only changes shape by the disappearance of a shoulder at lower values. This however causes the average R_{ee}, presented in Table 2, to increase. An increase of R_{ee} upon phosphorylation of Tau1 has been detected by fluorescence resonance energy transfer measurements, as reported by Chin et al. [15].

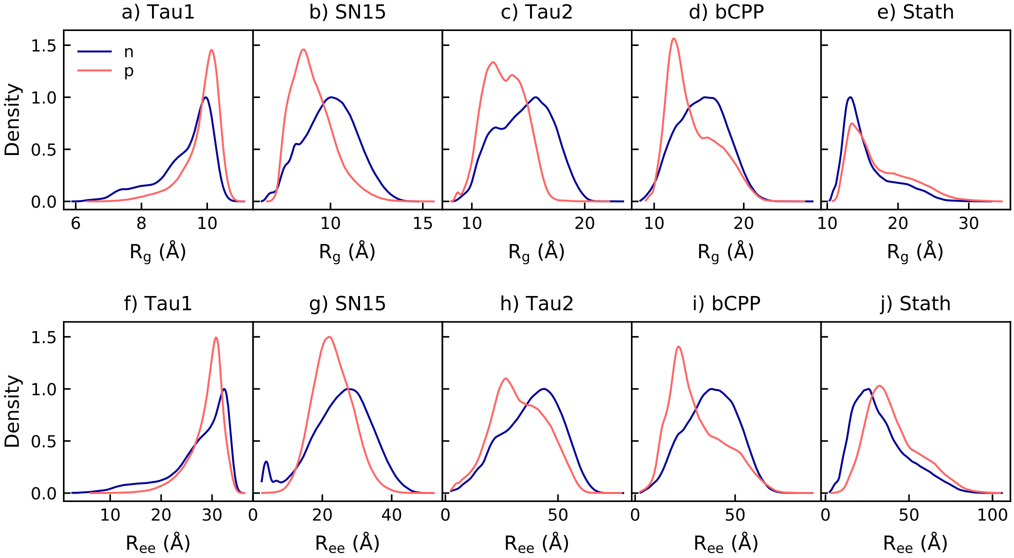


Figure 1. Radius of gyration (R_g) and end-to-end distance (R_{ee}) density distributions of the non-phosphorylated (n) and phosphorylated (p) variants. The SN15 data are taken from ref. [20].

Table 2. Average radius of gyration (R_g) and end-to-end distance (R_{ee}) of the non-phosphorylated (n) and phosphorylated (p) variants.

	Radius of gyration (nm)		End-to-end distance (nm)	
Peptide	n	p	n	p
Tau1	0.93 ± 0.01	0.98 ± 0.01	2.74 ± 0.06	2.89 ± 0.02
SN15	1.00 ± 0.01	0.90 ± 0.01	2.54 ± 0.09	2.30 ± 0.03
Tau2	1.46 ± 0.02	1.29 ± 0.03	3.83 ± 0.09	3.27 ± 0.17
bCPP	1.53 ± 0.03	1.43 ± 0.03	3.80 ± 0.08	3.09 ± 0.15
Stath	1.56 ± 0.04	1.73 ± 0.09	3.30 ± 0.24	4.05 ± 0.17

The shape factor, presented in Figure 2 can be used as an estimate of the shape of the peptide. If it behaves as a Gaussian coil, the shape factor is approximately 6, whereas for a stiff rod, it is around 12. SN15, Tau2, and bCPP are shown to behave rather coillike in non-phosphorylated state, while Tau1 is more stiff, and Stath more contracted. Upon phosphorylation bCPP becomes more contracted than a Gaussian coil, while Stath expands to become more coillike.

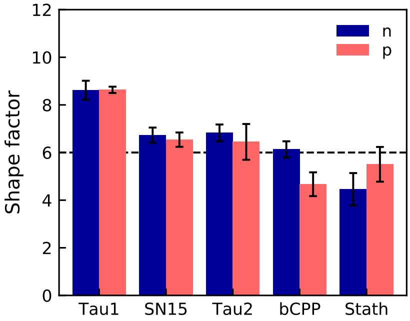


Figure 2. The shape factor of the non-phosphorylated (n) and phosphorylated (p) variants. The dashed line corresponds to the shape factor of a Gaussian coil. The error bars are based on error propagation of the error estimates determined for R_g and R_{ee} by block averaging.

Comparing the induced changes of R_g and R_{ee} with the net charge of the non-phosphorylated peptides, it is clear that the prediction of Jin and Gräter, i.e., that net charge controls the effect of phosphorylation [10], only holds for SN15, Tau2, and Stath. bCPP contracts despite having a negative net charge, and Tau1 expands despite the positive net charge. Hence, to understand the effect of phosphorylation of these peptides we need to investigate changes in secondary structure and specific interactions.

Phosphorylation of Tau1 favours expanded conformations

The average secondary structure content of the non-phosphorylated and phosphorylated variants of the peptides are shown in Figure 3. First, please notice that these peptides are all intrinsically disordered, as they are dominated by irregular structure. Several of the peptides are also shown to contain a substantial amount of polyproline type II helix (PPII), especially Tau1, which possess 46% and 51% PPII in the non-phosphorylated and phosphorylated state, respectively. Elam et al. [41] have predicted close to 50% PPII content in this region of Tau, and CD measurements of this segment indicate an increase of PPII content upon phosphorylation [15]. In Figure 4 it is shown that all structural changes upon phosphorylation at T175 and T81 take place at the C-terminal end of the peptide, from residue 179 and forward. The propensity for bends and turns at residue 179–181 decreases, while the PPII content increases at residue 181–182. There is occasional salt bridge formation between the phosphothreonines and their respective neighbouring lysine. Specifically, the probability of salt bridge formation is 7 ± 2% for pT175–K174 and 9 ± 2% for pT71–K180. The most occurring salt bridge is however formed between pT175 and the N-terminal, with a probability of 49 ± 9%. However, due to the close proximity between the salt bridging residues, the effect on the overall dimensions of the peptide is small. The conformational effects of phosphorylation of Tau is well summarized by Figure 5, showing the energy landscape and conformations. The energy landscape of non-phosphorylated Tau1 contains several minima, of which the minimum containing expanded conformations dominate, in line with the relatively high shape factor. Other less populated minima contain conformations with a kink in the C-terminal, originating from a bend or turn. Upon phosphorylation, the minima with kinked conformations disappears, leaving only the minima with expanded conformations. This explains the change in shape of the R_g and R_{ee} distributions, from a peak with preceding shoulder to a single peak.

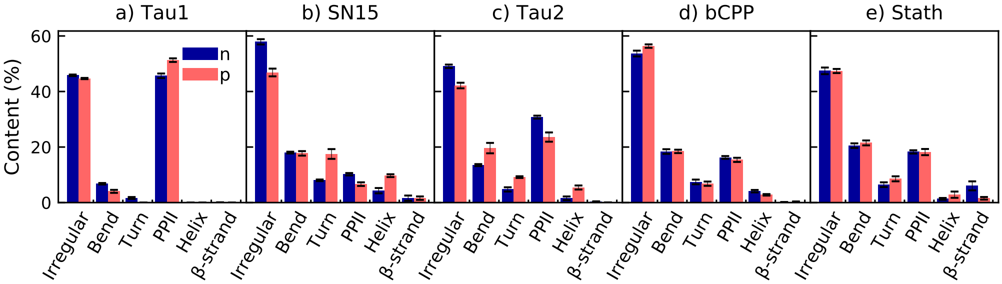


Figure 3. Average secondary structure content of the non-phosphorylated (n) and phosphorylated (p) variants. Helix includes α-helix and 3_{10}-helix. β-strand includes also β-bridge.

Phosphorylation increases the helix propensity and induces salt bridge formation in SN15 and Tau2

SN15, Tau2, and Stath all report an increase of helicity upon phopshorylation. The helical region, shown in Figure 4, corresponds to "pSpSEEKFLR" in SN15 and Stath, and "pSpSAKSR" in Tau2. The sequences, hence, share two characteristics: 1) the helical region starts with two phoshorylation sites, and 2) three or four steps away from the phosphorylation site is a positively charged residue positioned. Phosphorylation has been shown to stabilize α-helices if the phosphorylation site is located in the N-terminal end of the helix, by electrostatic interaction between phosphorylated serines and the macrodipole of the helix, and by hydrogen bonding with the amide backbone [42]. With a $i, i+4$ spacing between a phosphorylated serine and a lysine, phosphorylation also stabilizes α-helices through salt bridge formation between the side groups [43].

For Tau2 a phosphorylation-induced increase of α-helical structure from 5 to 40% in region A239–R242 has been reported [13]. In these simulations the main helical increase upon phosphorylation is associated with region S237–K240, where the increase is from 4 to 26%. However, the helical increase is mainly due to 3_{10}-helix, since the increase of α-helix is only from 1 to 5%. Hence, the simulations are in qualitative agreement with the experiments, but the quantitative results should be treated with caution. Also in SN15 the larger part of the helical increase is due to 3_{10}-helix, and an increase of α-helix is supported by CD spectroscopy [20], once again giving qualitative support to the findings in this study. Notice also that while it is hard to make quantitative comparisons with CD data, our study on SN15 suggested that the simulations underestimate the structural content [20], which is the same as observed for Tau2.

While helix formation decreases the R_g and R_{ee} also salt bridge formation can contribute to the compaction observed upon phosphorylation. In SN15 the salt bridges pS2–K6, pS3–K6, pS3–R9, and pS3–R10 are the most probable and all form with an approximately 25% occurrence. From the contact map in Supplementary Figure S1, it appears that the pS2–K6 and pS3–K6 salt bridges contribute to stabilize the formed helix. The pS3–R9 and pS3–R10 salt bridges are also visible in the contact map and contribute to an increase in the amount of more compact conformations. In the energy landscape in Figure 6, it is shown that phosphorylation shifts the position of the main minima in the energy landscape, from an area of more coil-like structures to a more compact state. The non-phosphorylated peptide also samples conformations that are more compact with a higher content of sec-

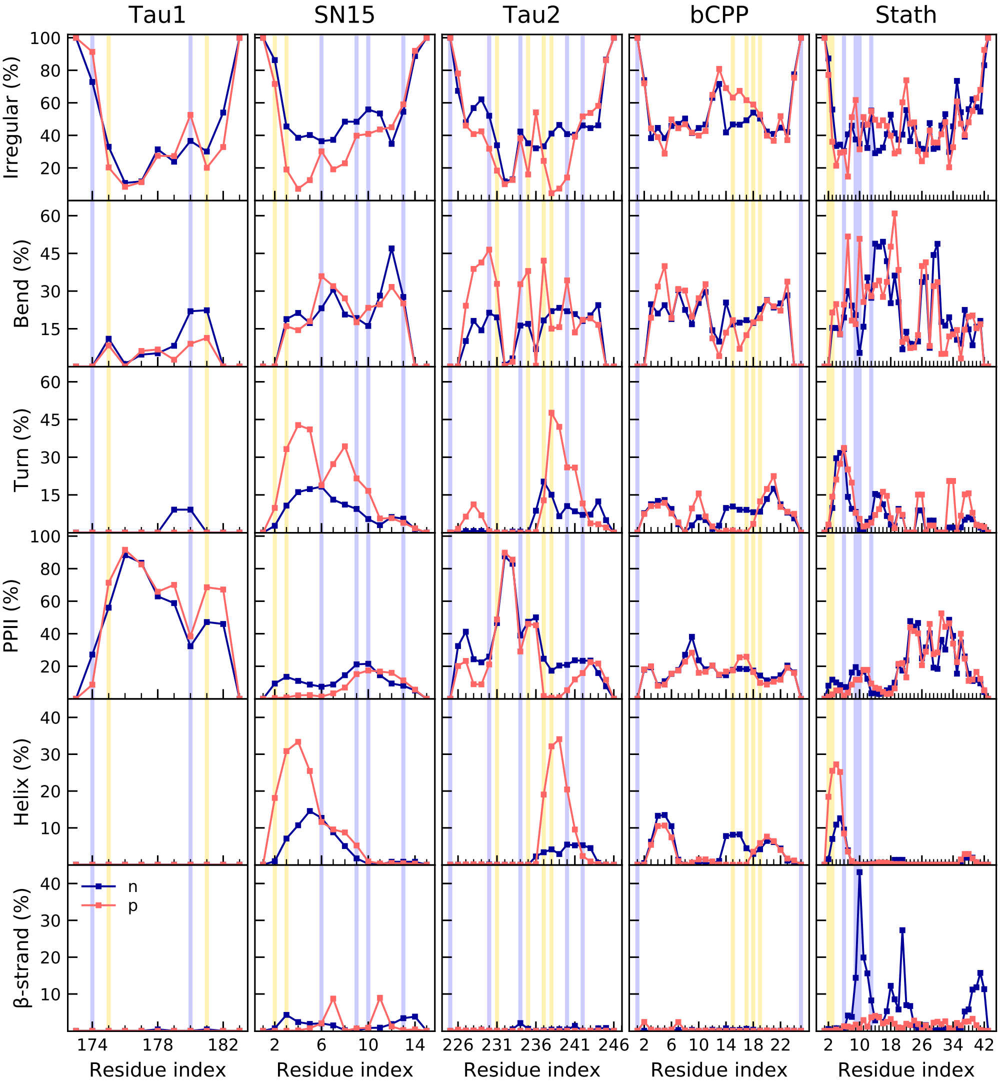


Figure 4. Secondary structure content along the non-phosphorylated (n) and phosphorylated (p) sequence. Helix includes α-helix and 3_{10}-helix. β-strand includes also β-bridge. The position of phosphorylated and positively charged residues are highlighted in yellow and blue, respectively. The SN15 data are taken from ref. [20].

ondary structure, but more rarely than the phosphorylated peptide. The conformation corresponding to the minimum in the most populated basin in the phosphorylated peptide have residue pS2 and K6 close enough to be in contact, however, there is no helix, but instead a turn at residues E4–E5. This shows that it is favourable to have pS2 and K6 in

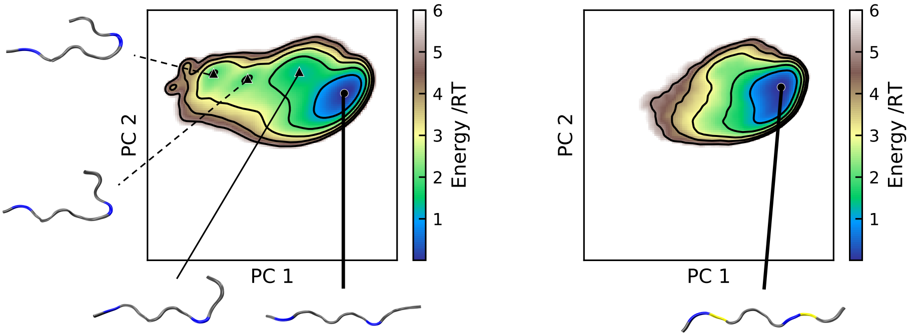


Figure 5. Energy landscapes and conformations in minima of Tau1. Left: non-phosphorylated, right: phosphorylated. The energy landscapes are constructed using the first two components from principal component analysis, applying the same basis set for both variants. Hence, they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$. A thick line corresponds to the most populated basin, while dashed lines to the least populated basins. In the conformations positively charged residues are shown in blue, and phosphorylated residues in yellow.

contact, but that the interaction does not necessarily imply helix formation. In Figure 4, it was shown that also the turn content in region S3–E5 increases upon phosphorylation, not only the helix content. There is also an increase of turn content in region F7–R11, which is partly caused by occasional β-strand formation, as shown in the other conformation in Figure 6, and partly by residues pS3 and R9 coming close to form a salt bridge. Both of these changes give rise to more compact conformations. We must however note that SAXS measurements have indicated that a compaction upon phosphorylation is plausible, but probably smaller than shown in the simulations [20]. While Jin and Gräter found that changes in the hydration shell upon phosphorylation can hide global conformational changes in SAXS measurements, they also concluded that the force field used in this study overestimates the charge effect, thus providing two different explanations of the deviations between the simulations and experiments [10].

In Tau2 several salt bridges have been established from NMR measurements, specifically pT231–R230, pS237–K240, and pS238–R242 [13]. pT231–R230 and pS238–R242 are indeed the two most occurring salt bridges according to Table 3, while pS237–R242 is the third most common. Apart from the increase of helical content related to phosphorylation, Figure 4 reveals an interesting pattern of bends after phosphorylation, where the charged residues R, K, pT, and pS are enriched in bends. The conformations in Figure 7 illustrate how the salt bridges contribute to the formation of bends. Since the probability of a turn at A227-V229 is roughly the same as the probability of the pT231–K225 salt bridge (see Figure 4 and 3), and V228 is located right between K225 and pT231, we conclude that also this turn is a result of a salt bridge interaction. Hence, also this peptide show that salt bridge formation induces bends and turns.

Comparing the energy landscapes of non-phosphorylated and phosphorylated Tau2 in Figure 7, it is shown that for both peptides more extended conformations, such as in the minima furthest to the right, are sampled, but to different extent. These type of conformations are more common in the non-phosphorylated variant, while the most populated basin contains conformations with the N-terminal end folded over, to come closer to the phosphorylated residues. While K225 rarely involves in a proper salt bridge with other residues than pT231, it is still energetically favourable to be in rather close vicinity of the phosphorylated region, considering both the charged side chain and the N-

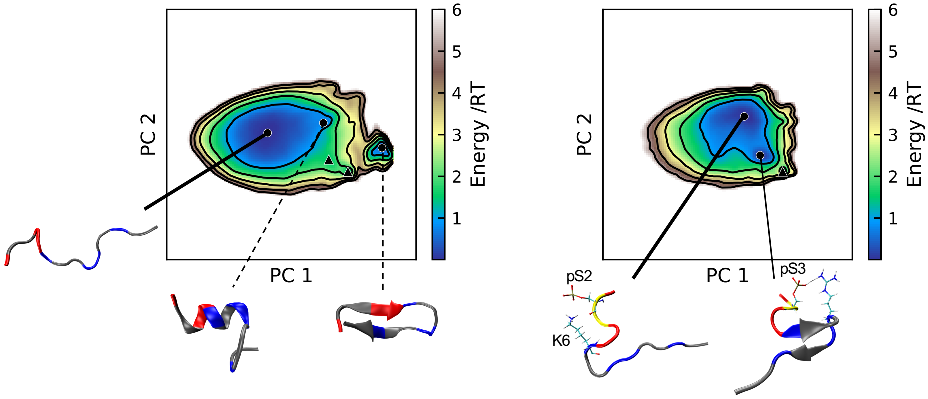


Figure 6. Energy landscapes and conformations in the lowest energy minima of SN15. Left: non-phosphorylated, right: phosphorylated. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both variants. Hence, they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$. A thick line corresponds to the most populated basin, while dashed lines to the least populated basins. In the conformations, positively charged residues are shown in blue, negatively charged residues in red, and phosphorylated residues in yellow. Phosphorylated and positively charged residues that are close are shown explicitly.

Table 3. Probability of salt bridge formation (%) between phosphorylated residues and positively charged residues in Tau2, where NT is the N-terminus. The values printed in bold corresponds to the experimentally established salt bridges [13].

Residue	NT	K225	R230	K234	K240	R242
pT231	1 ± 1	10 ± 3	**37 ± 10**	3 ± 2	∼ 0	∼ 0
pS235	< 1	2 ± 1	< 1	15 ± 4	17 ± 2	6 ± 3
pS237	2 ± 1	4 ± 3	3 ± 10	17 ± 2	**19 ± 2**	29 ± 2
pS238	4 ± 1	5 ± 2	3 < 1	∼ 0	5 ± 4	**35 ± 6**

terminus. This type of conformations give rise to an increased contact probability within the N-terminal part of the chain, see Supplemental Figure S2. Ignoring the contacts close to the diagonal, which indicates helical structure and certain salt bridges, the non-phosphorylated variant has a higher probability of contacts within the C-terminal end. The two minima in the left part of the energy landscape in Figure 7 are examples of such conformations, which originates from the electrostatic attraction between the C-terminus and the positively charged residues. Notice however, that the probability of conformations with one end folded over is much higher after phosphorylation, which explains the decrease in R_g and R_{ee}. The conformation corresponding to the minimum in the most populated basin for the phosphorylated peptide additionally shows a helix in the C-terminal end, which also contributes to a decreased R_g and R_{ee}.

Salt bridge formation shifts the conformational ensemble of bCPP

For bCPP the secondary structure content is highly similar in phosphorylated and non-phosphorylated state, as shown by Figure 3, in agreement with CD spectroscopy results by Farrell et al. [25]. The small difference that occurs upon phosphorylation at S14, S17, S18, and S19 is a change from helix and turn to irregular structure in region E14–S17, see Figure

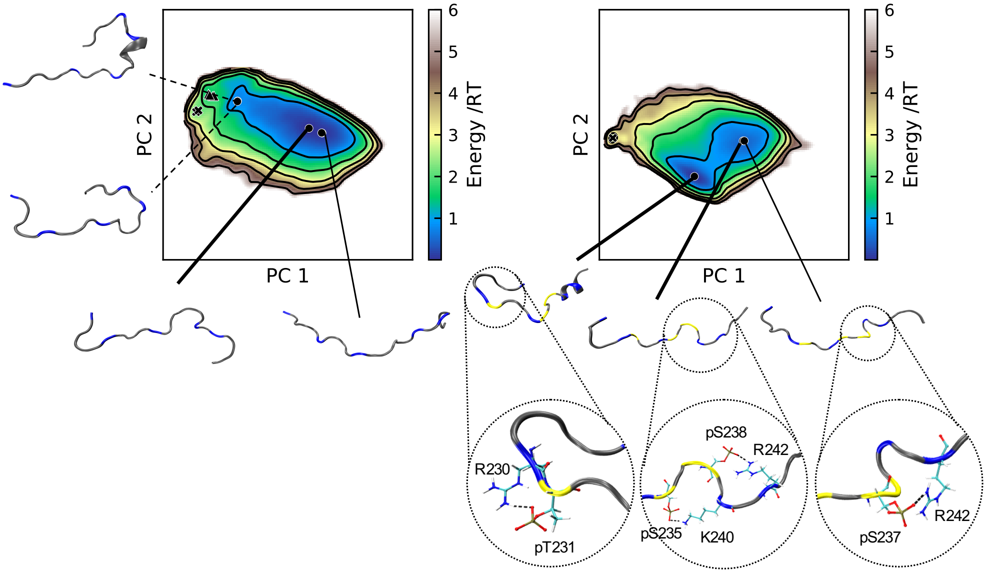


Figure 7. Energy landscapes and conformations in the lowest energy minima of Tau2. Left: non-phosphorylated, right: phosphorylated. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both variants, hence making them directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$, ✖: $\leq 3RT$. Thick lines correspond to the most populated basins, while dashed lines to the least populated basins. In the conformations positively charged residues are shown in blue, and phosphorylated residues in yellow.

4. The vanishing of helical content is in agreement with the conclusion of Andrew et al., that phosphorylation of a residue in the interior of a helix, without a positively charged residue within suitable distance, destabilizes the helix [42]. Since disruption of a short helix would not cause a contraction of the peptide, the conformational changes in bCPP upon phosphorylation is not explained by secondary structure. Instead, the contraction is due to electrostatic attraction including salt bridge formation between the positively charged end residues and the phosphorylated residues, as seen in Table 4. Although both end residues are arginines, there is preference of R1 to interact with the phosphorylated region over R25, due to the respective charges of the terminii. This is evident from the fact that also the N-terminus involves in salt bridges with the phosphorylated residues, and further shown in the contact map in Supplementary Figure S3. When R1 interacts with the phosphorylated residues, it causes the peptide to fold over, reducing R_g and R_{ee} substantially. From the energy landscape in Figure 8, it is shown that before phosphorylation the minima with lowest energy contain more extended conformations, while after phosphorylation the minima with lowest energy instead showcase the N-terminal part being folded over.

Based only on the net charge of non-phosphorylated bCPP, it was expected that it would expand upon phosphorylation. Considering only region E13–E21, which contains the four phosphorylation sites, this effect was noticed. The average distance between the C_α atoms of residue 13 and 21 increases from 1.91 ± 0.03 nm to 2.12 ± 0.03 nm upon phosphorylation. However, due to the strong electrostatic interaction between the arginines and the phosphorylated region that are far apart in the sequence, the global result is compaction. Hence, the relative position of charged residues is very important to consider for the effects of phosphorylation on the overall dimensions of the peptide.

Table 4. Probability of salt bridge formation (%) between phosphorylated residues and positively charged residues in bCPP, where NT is the N-terminus.

Residue	NT	R1	R25
pS15	2 ± 1	6 ± 1	2 ± 1
pS17	3 ± 1	7 ± 1	7 ± 2
pS18	4 ± 1	13 ± 4	12 ± 4
pS19	1 ± 1	10 ± 4	15 ± 4

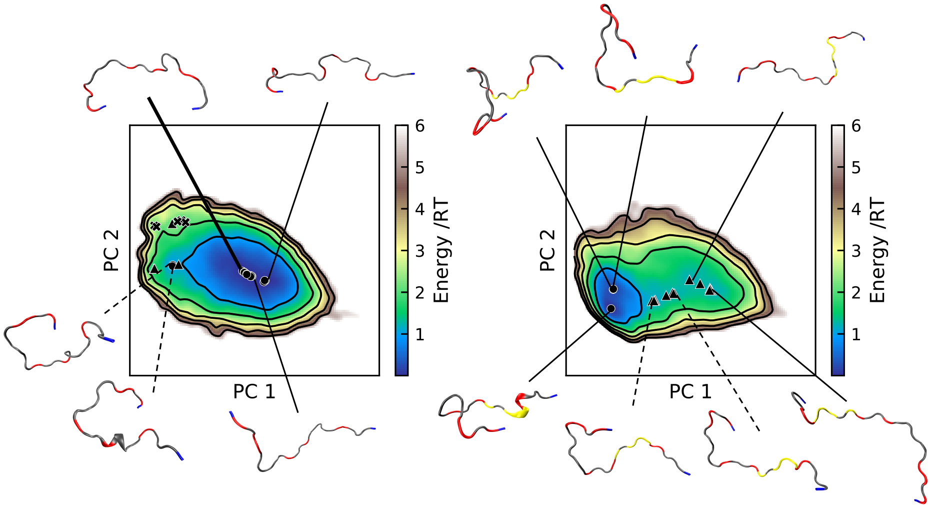


Figure 8. Energy landscapes and conformations in selected minima of bCPP. Left: non-phosphorylated, right: phosphorylated. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both variants. Hence, they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$, ✖: $\leq 3RT$. A thick line corresponds to the most populated basin, while dashed lines to the least populated basins. In the conformations positively charged residues are shown in blue, negatively charged residues in red and phosphorylated residues in yellow.

We previously showed that addition of 150 mM NaCl had negligible effects on the salt bridges and global conformational properties of phosphorylated bCPP [31]. The same applies to non-phosphorylated bCPP, as presented in Supplementary Figure S4–S5. However, although the average values of R_g at 0 and 150 mM are within error, there is a slight increase in the phosphorylated variant and decrease in the non-phosphorylated variant, see Table 5. Hence, at 150 mM NaCl, the difference observed in R_g between the two variants vanishes, considering the associated error. Note however that the distributions still have distinctly different shapes, hence we argue that the conformational ensembles are still different. The same trend is observed in the average R_{ee} values, although a difference with respect to phosphorylation state still remains at 150 mM NaCl, see Table 5. Also in the calculated scattering curve (Supplementary Figure S5) is the effect of salt smaller than the effect of phosphorylation. The difference between the form factor of non-phosphorylated and phosphorylated bCPP is however still rather small, so we suspect that it can be hard to detect experimentally with SAXS. Based on the fraction of charged residues and level of charge separation, we expect the other peptides in this study to show smaller effects in regards to salt concentration than bCPP. Hence, we expect the results observed here to be valid also at 150 mM NaCl.

Table 5. Average radius of gyration and end-to-end distance of the non-phosphorylated (n) and phosphorylated (p) bCPP in the presence of 0 and 150 mM NaCl.

	Radius of gyration (nm)		End-to-end distance (nm)	
	0 mM	150 mM	0 mM	150 mM
n	1.53 ± 0.03	1.48 ± 0.02	3.80 ± 0.08	3.64 ± 0.09
p	1.43 ± 0.03	1.45 ± 0.03	3.09 ± 0.15	3.37 ± 0.13

Arginine–phosphoserine interactions outshines arginine–tyrosine interactions in Stath

Upon phosphorylation of Stath, the three largest changes in secondary structure is a decrease of β-strand structure, an increase of helical structure, and an increase of turns. Figure 4 implies that residues R10, Y18, Y21, and Y41 are of extra importance for the formation of β-sheet. The cation-π interaction that can occur between aromatic residues, such as tyrosine, and cationic residues, such as arginine, have been shown to be common in proteins [44]. A correlation between β-strands and cation-π interactions have also been established [45]. Table 6 show that the cation-π interaction indeed is more occurring in non-phosphorylated Stath than in phosphorylated Stath, suggesting that it drives the formation of β-strands. The conformations in Figure 9i-iii show examples of the cation-π interaction in non-phosphorylated Stath. Although the aromatic–cation interactions are more common in non-phosphorylated Stath, they still occur in phosphorylated Stath, as exemplified by Figure 9. Upon phosphorylation the occurrence of cation–π interaction decreases substantially, while salt bridge formation appears according to Table 7. Notice that R10, which was shown to interact with tyrosines, is involved in one of the most probable salt bridges, pS3–R10. Hence, the arginine–phosphoserine interaction is deemed stronger than the arginine–tyrosine interaction. The replacement of arginine–tyrosine interaction with arginine–phosphoserine causes the β-strands to vanish, which explains the observed expansion.

Table 6. Probability of cation–π interaction (%) for certain pairs of residues in non-phosphorylated (n) and phosphorylated (p) Stath.

Residues	n	p
R10–Y18	13.8 ± 6.3	1.6 ± 0.9
R10–Y21	32.0 ± 8.6	3.9 ± 0.7
R10–Y41	9.2 ± 4.3	0.4 ± 0.2

Table 7. Probability of salt bridge formation (%) between phosphorylated residues and positively charged residues in Stath, where NT is the N-terminus.

Residue	NT	K6	R9	R10	R13
pS2	< 1	23 ± 7	23 ± 8	12 ± 1	8 ± 1
pS3	12 ± 3	9 ± 1	30 ± 8	32 ± 7	6 ± 3

As presented above, SN15, which is the first fifteen residues of Stath, contracts upon phosphorylation, which was explained by the increased helicity and formation of salt bridges. Supplementary Figure S6 shows that in phosphorylated Stath, the global dimensions of the first fifteen residues, Stath$_{1\text{–}15}$ agree with those of the fragment (SN15). In the

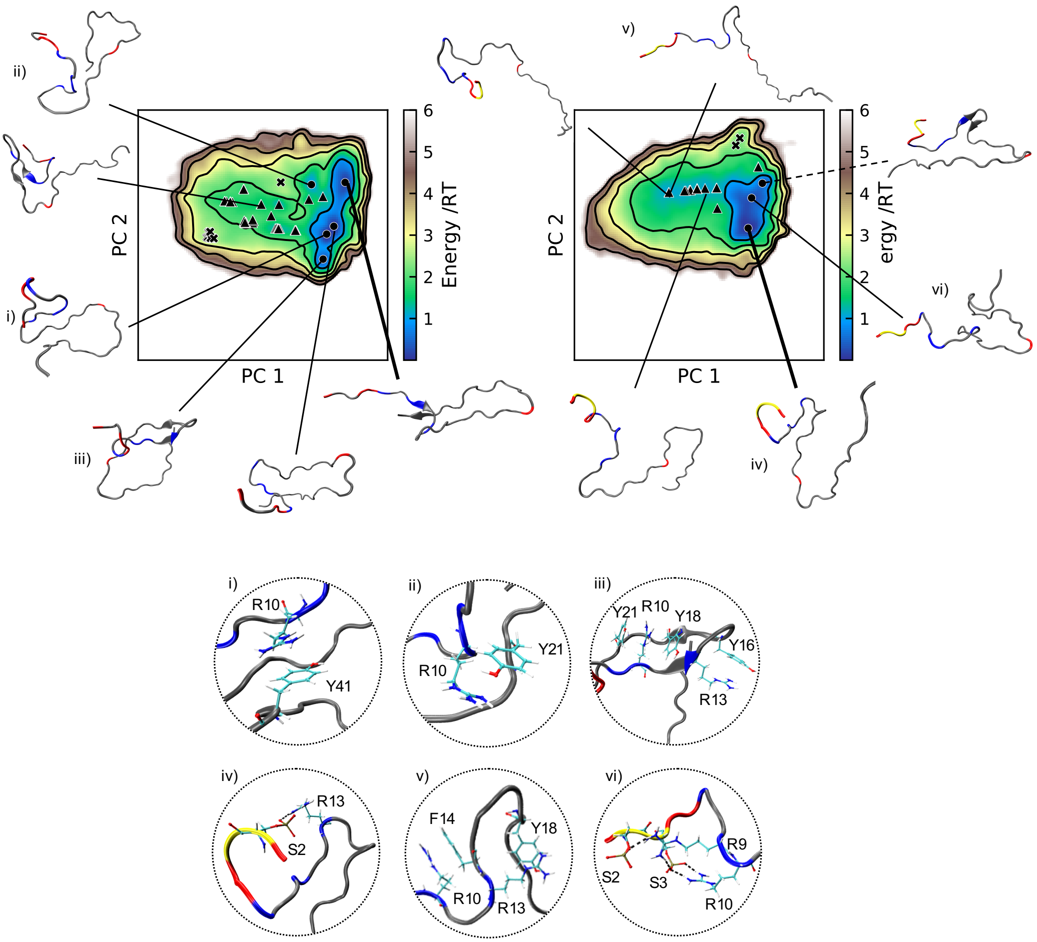


Figure 9. Energy landscapes and conformations in the minima of the most populated basins of Stath. Left: non-phosphorylated, right: phosphorylated. The energy landscapes are constructed using the first two components from principal component analysis, using the same basis set for both variants, hence making them directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$, ✖: $\leq 3RT$. A thick line corresponds to the most populated basin, while dashed line to the least populated basin. In the conformations positively charged residues are shown in blue, negatively charged residues in red and phosphorylated residues in yellow. The circles show specific interactions within the peptide in the conformations corresponding to the letters.

non-phosphorylated variant the distributions are also rather similar, except for a sharp peak in both the R_g and R_{ee} distributions, which corresponds to a basin in the energy landscape with the conformation shown in Supplementary Figure S6c. Regarding the secondary structure, according to Supplementary Figure S7, the largest difference between SN15 and $Stath_{1-15}$ is caused by β-strand not forming in SN15, due to lacking its partner further on in the sequence. There are also some differences in bends and turns, but the increase of helical propensity is similar. Hence, overall, the first fifteen residues of Stath behaves rather similarly in the full peptide and as a standalone fragment, although especially the presence of the rest of the sequence induces β-strand formation. Despite this discrepancy, we can conclude that phosphorylation of Stath causes a contraction of the first fifteen residues, but an expansion of the full peptide, due to disruption of β-sheets.

Conclusions

Some of the peptides in this study contracted upon phosphorylation, while others became more expanded. However, the net charge was not enough to predict the effect. Instead, we have identified factors that appeared to be of greater importance, of which the first is the distribution of charged residues, in line with the influence of linear charge distribution on the conformational ensemble of IDPs [46]. Especially the relative position of phosphorylated and positively charged residues mattered, considering that salt bridges formed between residues far from each other in the sequence had the largest effect on the overall dimensions of the peptide. Regarding salt bridges, Kumar et. al have shown that phosphorylation can re-wire salt bridges by competing with already present E–R salt bridges [47], but no such tendencies were observed for these peptides. Here the possible salt bridges in the non-phosphorylated peptides were either low in probability or did not change much upon phosphorylation. In Stath, competitive interactions between positively charged residues, aromatic residues, and phosphorylated residues accounted for the changes upon phosphorylation. This shows that for peptides which include arginine, it can be of importance to also consider aromatic residues. In both bCPP and Stath, phosphorylation induced the opposite effect on the local and global dimensions, hence, to understand the purpose/implications of the phosphorylated residues, both length-scales should be studied. This is especially important dealing with longer IDPs where local/non-local effects can have larger compensatory effect than observed for short peptides [14].

Regarding secondary structure, the separation between phosphorylated and positively charged residues were shown to control the helix propensity, and salt bridges additionally induced changes in the amount of bends and turns. Comparison with experimental data on secondary structure for SN15 and Tau2 indicates that the simulations underestimate the structural content. For these peptides a preference of 3_{10}- over α-helix was also observed, while the experimental data only considered α-helix. Hence, the simulations were better at indicating trends than produce exact measurements of secondary structure. Overall, the simulation results were often in qualitative agreement with available experimental data, suggesting that despite the deficiency related to secondary structure and the reported tendency of the force field to overestimate charge–charge interactions, simulations with this force field can still contribute to an increased understanding of the implications of phosphorylation.

As a final note, this study shows that there are several factors contributing to the outcome of phosphorylation, and that they are of varying importance in different peptides. This shows that phosphorylation indeed is complex, however, it is still possible to obtain a better understanding of these factors individually. Therefore, we have an ongoing project in which the number of phosphorylated residues and their positions are varied in a controlled manner, to investigate the effects of those factors systematically.

Materials and Methods

All-atom molecular dynamics simulations of the systems shown in Table 8 were performed using GROMACS version 2018.4 (version 4.6.7 for simulation of Stathn) [48–52] with the AMBER ff99SB-ILDN [53] force field and the TIP4P-D [54] water model. Parameters for phosphorylated residues were derived from Homeyer et al. [55] and Steinbrecher et al. [56].

Initial configurations of the peptides were constructed from the sequence as linear chains using Avogadro 1.2.0 [57], optimizing the structure with the auto-optimization tool. SN15n and Stathn were constructed as linear chains in PyMOL [58]. Each peptide was

Table 8. Details of the simulations performed in this work. The suffix n stands for non-phosphorylated peptide, while the suffix p stands for phosphorylated.

Peptide	Box volume (nm^3)	Number of solvent molecules	Number of sodium ions	Number of chloride ions	Total simulation length (µs)
Tau1n	157.63	5130	0	2	10.0
Tau1p	140.55	4594	2	0	5.0
Tau2n	724.974	23862	0	5	6.0
SN15n[a]	272.13	8839	0	1	14.4
SN15p[a]	294.52	9703	3	0	22.0
Tau2p[b]	722.941	23816	3	0	11.0
bCPPn	1009.24	32975	5	0	5.0
bCPPn, 150 mM	1009.24	32793	96	91	5.0
bCPPp[b]	1002.41	32815	13	0	6.0
bCPPp, 150 mM[b]	1002.41	32633	104	91	7.0
Stathn [c]	930.47	30651	0	0	17.0
Stathp[b]	942.11	30942	4	0	12.0

[a] Previously published [20].
[b] Accepted for publication [31].
[c] Using GROMACS version 4.6.7.

placed in a rhombic dodecahedron box with a minimum distance between the peptide and the box edges of 10 Å, and solvated. The number of water molecules is specified in Table 8, alongside the number of chloride and sodium ions that were added to neutralize the system and in two cases obtain a salt concentration of 150 mM. Periodic boundary conditions were employed in all directions. The equations of motion were integrated using the Verlet leapfrog algorithm [59] with a time step of 2 fs. Non-bonded interactions were treated with a Verlet list cutoff scheme. The short-ranged interactions were calculated using neighbour lists with cutoff 10 Å. Long-ranged dispersion corrections were applied to energy and pressure and long-ranged electrostatic interactions were treated by Particle Mesh Ewald [60] with a cubic interpolation and 1.6 Å grid spacing. All bond lengths were constrained using the LINCS algorithm [61]. Solute and solvent were separately coupled to temperature baths at 298 K using the velocity rescaling thermostat [62] with a 0.1 ps relaxation time. Parrinello-Raman pressure coupling [63] was used to keep the pressure at 1 bar, using a 2 ps relaxation time and $4.5 \cdot 10^{-5}$ bar^{-1} isothermal compressibility.

Energy minimization was performed by the steepest descent algorithm until the system was converged within the available machine precision. Initiation of five replicates per system with different starting seeds were performed separately in two steps using position restraints on the peptide. The first step was 500 ps of NVT simulation (constant number of particles, volume, and temperature) performed to stabilize the temperature, followed by the second step of 1000 ps of NPT simulation (constant number of particles, pressure, and temperature) to stabilize the pressure. Production runs of the five replicates per system were performed in the NPT ensemble, for at least 1 µs per replicate. bCPPp with 150 mM salt was simulated in 10 replicates for 0.7 µs each. The total simulation time per system is stated in Table 8. Energies and coordinates were saved every 10 ps, except for in the simulations with 150 mM NaCl. There the saving frequency was every 50 or 40 ps, for bCPPn and bCPPp, respectively.

Analysis

R_g and R_{ee} were calculated using GROMACS 2018.4 and the *gmx analyze* routine was used to obtain averages and error estimates from block averaging analysis. Distributions were obtained by Gaussian kernel estimation using the SciPy package version 1.5.4 [64]. The shape factor, r_s, was calculated from the average values of R_g and R_{ee} according to:

$$r_s = \frac{R_{\mathrm{ee}}^2}{R_{\mathrm{g}}^2}. \tag{1}$$

Secondary structure was determined using the DSSP program version 2.2.1 [65] with an extension to detect polyproline type II structure [66,67], on 10 000 equally spaced frames from the combined trajectory. The MDTraj Python library version 1.9.3 [68] was used to obtain contact maps, analyze salt bridges, and cation–π interactions. Since salt bridges are formed as a result of hydrogen bonding and electrostatic interactions, they have been assessed by analyzing the presence of hydrogen bonds based on the criterion in reference [69], as implemented in MDTraj. Cation–π interactions were analyzed based on the position of the NZ atom in arginine and CG and CZ in tyrosine. Interaction was defined to occur when both the distances R:NZ–Y:CG and R:NZ–Y:CZ were $\leq$ 6 Å [44]. The energy landscapes were calculated using principal component analysis following the approach described by Campos and Baptista [70], with the differences described by Henriques et al. [71]. In short, principal component analysis was applied to the cartesian coordinates of the backbone atoms of the protein, obtained after translational and rotational least square fitting on the central structure of the simulation. The conditional free energy was calculated from the probability density function in the representation space constructed by the first two principal components, obtained by Gaussian kernel density estimation. Snapshots from the simulations were produced using VMD 1.9.3 [72–74]. Data were plotted using Jupyter Notebook [75] with Python version 3.6.4 and packages NumPy version 1.19.5 [76] and Matplotlib version 2.1.2 [77].

Convergence and sampling quality were assessed by comparing the R_g and R_{ee} distributions, and energy landscapes, between the replicates, as well as by observing the auto-correlation function and convergence of the block average error estimate of R_g and R_{ee} in the concatenated simulation. These data are available in Supplementary File S2.

Supplementary Materials: Supplementary File S1 contains supplementary figures of contact maps, salt effects, and SN15 versus $\text{Stath}_{1\text{-}15}$. Supplementary File S2 contains figures for sampling and convergence assessment of the simulations.

Funding: Financial support has been given by the Royal Physiographic Society in Lund and the Bertil Andersson's foundation.

Acknowledgments: The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Center for Scientific and Technical Computing at Lund University (LUNARC).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IDP	Intrinsically disordered protein
SAXS	Small-angle X-ray scattering
CD	Circular Dichroism
NMR	Nuclear Magnetic Resonance
R_g	Radius of gyration
R_{ee}	End-to-end distance

References

1. Dunker, A.; Lawson, J.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; Ausio, J.; Nissen, M.S.; Reeves, R.; Kang, C.; Kissinger, C.R.; Bailey, R.W.; Griswold, M.D.; Chiu, W.; Garner, E.C.; Obradovic, Z. Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001**, *19*, 26–59. doi:https://doi.org/10.1016/S1093-3263(00)00138-8.

2. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533. doi:https://doi.org/10.1016/S0968-0004(02)02169-2.
3. Fisher, C.K.; Stultz, C.M. Constructing ensembles for intrinsically disordered proteins. *Current Opinion in Structural Biology* **2011**, *21*, 426–431. doi:https://doi.org/10.1016/j.sbi.2011.04.001.
4. Uversky, V.N. Wrecked regulation of intrinsically disordered proteins in diseases: pathogenicity of deregulated regulators. *Frontiers in Molecular Biosciences* **2014**, *1*, 6. doi:10.3389/fmolb.2014.00006.
5. 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. doi:https://doi.org/10.1016/j.sbi.2011.03.011.
6. Dunker, A.K.; Brown, C.J.; Lawson, J.D.; Iakoucheva, L.M.; Obradovićá, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **2002**, *41*, 6573–6582, [https://doi.org/10.1021/bi012159+]. PMID: 12022860, doi:10.1021/bi012159+.
7. Iakoucheva, L.M.; Radivojac, P.; Brown, C.J.; O'Connor, T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004**, *32*, 1037–1049, [http://oup.prod.sis.lan/nar/article-pdf/32/3/1037/9489921/gkh253.pdf]. doi:10.1093/nar/gkh253.
8. Gao, J.; Xu, D., Correlation Between Posttranslational Modification and Intrinsic Disorder in Protein. In *Biocomputing 2012*; Altman, R.B.; Dunker, A.K.; Hunter, L.; Murray, T.A.; Klein, T.E., Eds.; World Scientific Publishing Co. Pte. Ltd.; pp. 94–103. doi:10.1142/9789814366496_0010.
9. Johnson, L.N.; Lewis, R.J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **2001**, *101*, 2209–2242, [https://doi.org/10.1021/cr0002 PMID: 11749371, doi:10.1021/cr000225s.
10. F, J.; F, G. How multisite phosphorylation impacts the conformations of intrinsically disordered proteins. *PLoS Comput Biol* **2021**, *17*, e1008939. doi:https://doi.org/10.1371/journal.pcbi.1008939.
11. Firman, T.; Ghosh, K. Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. *The Journal of Chemical Physics* **2018**, *148*, 123305, [https://doi.org/10.1063/1.5005821]. doi:10.1063/1.5005821.
12. Uversky, V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Frontiers in Physics* **2019**, *7*, 10. doi:10.3389/fphy.2019.00010.
13. Schwalbe, M.; Kadavath, H.; Biernat, J.; Ozenne, V.; Blackledge, M.; Mandelkow, E.; Zweckstetter, M. Structural Impact of Tau Phosphorylation at Threonine 231. *Structure* **2015**, *23*, 1448–1458. doi:https://doi.org/10.1016/j.str.2015.06.002.
14. Martin, E.W.; Holehouse, A.S.; Grace, C.R.; Hughes, A.; Pappu, R.V.; Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *Journal of the American Chemical Society* **2016**, *138*, 15323–15335, [https://doi.org/10.1021/jacs.6b10272]. PMID: 27807972, doi:10.1021/jacs.6b10272.
15. Chin, A.F.; Toptygin, D.; Elam, W.A.; Schrank, T.P.; Hilser, V.J. Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. *Biophysical Journal* **2016**, *110*, 362–371. doi:https://doi.org/10.1016/j.bpj.2015.12.013.
16. Kulkarni, P.; Jolly, M.K.; Jia, D.; Mooney, S.M.; Bhargava, A.; Kagohara, L.T.; Chen, Y.; Hao, P.; He, Y.; Veltri, R.W.; Grishaev, A.; Weninger, K.; Levine, H.; Orban, J. Phosphorylation-induced conformational dynamics in an intrinsically disordered protein and potential role in phenotypic heterogeneity. *Proceedings of the National Academy of Sciences* **2017**, *114*, E2644–E2653, [https://www.pnas.org/content/114/13/E2644.full.pdf]. doi:10.1073/pnas.1700082114.
17. Wang, K.; Ning, S.; Guo, Y.; Duan, M.; Yang, M. The regulation mechanism of phosphorylation and mutations in intrinsically disordered protein 4E-BP2. *Phys. Chem. Chem. Phys.* **2020**, *22*, 2938–2948. doi:10.1039/C9CP05888E.
18. Rani, L.; Mittal, J.; Mallajosyula, S.S. Effect of Phosphorylation and O-GlcNAcylation on Proline-Rich Domains of Tau. *The Journal of Physical Chemistry B* **2020**, *124*, 1909–1918, [https://doi.org/10.1021/acs.jpcb.9b11720]. PMID: 32065850, doi:10.1021/acs.jpcb.9b11720.
19. Liu, N.; Guo, Y.; Ning, S.; Duan, M. Phosphorylation regulates the binding of intrinsically disordered proteins via a flexible conformation selection mechanism. *Communications Chemistry* **2020**, *3*, 123. doi:10.1038/s42004-020-00370-5.
20. Rieloff, E.; Skepö, M. Phosphorylation of a Disordered Peptide—Structural Effects and Force Field Inconsistencies. *J. Chem. Theory Comput.* **2020**, *16*, 1924–1935, [https://doi.org/10.1021/acs.jctc.9b01190]. PMID: 32050065, doi:10.1021/acs.jctc.9b01190.
21. Willet, A.H.; Igarashi, M.G.; Chen, J.S.; Bhattacharjee, R.; Ren, L.; Cullati, S.N.; Elmore, Z.C.; Roberts-Galbraith, R.H.; Johnson, A.E.; Beckley, J.R.; Gould, K.L. Phosphorylation in the intrinsically disordered region of F-BAR protein Imp2 regulates its contractile ring recruitment. *Journal of Cell Science* **2021**, *134*, [https://journals.biologists.com/jcs/article-pdf/134/16/jcs258645/2100845/jcs258645.pdf]. jcs258645, doi:10.1242/jcs.258645.
22. Papamokos, G.V.; Tziatzos, G.; Papageorgiou, D.G.; Georgatos, S.; Kaxiras, E.; Politou, A.S. Progressive Phosphorylation Modulates the Self-Association of a Variably Modified Histone H3 Peptide. *Frontiers in Molecular Biosciences* **2021**, *8*, 558. doi:10.3389/fmolb.2021.698182.
23. Nicolaou, S.T.; Hebditch, M.; Jonathan, O.J.; Verma, C.S.; Warwicker, J. PhosIDP: a web tool to visualize the location of phosphorylation sites in disordered regions. *Scientific Reports* **2021**, *11*, 9930. doi:10.1038/s41598-021-88992-0.
24. Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J.D. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* **2010**, *18*, 494–506. doi:https://doi.org/10.1016/j.str.2010.01.020.
25. Farrell, H.; Qi, P.; Wickham, E.; Unruh, J. Secondary Structural Studies of Bovine Caseins: Structure and Temperature Dependence of β-Casein Phosphopeptide (1-25) as Analyzed by Circular Dichroism, FTIR Spectroscopy, and Analytical Ultracentrifugation. *J Protein Chem* **2002**, *21*, 307—321, [https://doi.org/10.1023/A:1019992900455]. doi:10.1023/A:1019992900455.

26. Brister, M.A.; Pandey, A.K.; Bielska, A.A.; Zondlo, N.J. OGlcNAcylation and Phosphorylation Have Opposing Structural Effects in tau: Phosphothreonine Induces Particular Conformational Order. *Journal of the American Chemical Society* **2014**, *136*, 3803–3816, [https://doi.org/10.1021/ja407156m]. PMID: 24559475, doi:10.1021/ja407156m.
27. Chong, S.H.; Chatterjee, P.; Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Annual Review of Physical Chemistry* **2017**, *68*, 117–134, [https://doi.org/10.1146/annurev-physchem-052516-050843]. PMID: 28226222, doi:10.1146/annurev-physchem-052516-050843.
28. Huang, J.; MacKerell, A.D. Force field development and simulations of intrinsically disordered proteins. *Current Opinion in Structural Biology* **2018**, *48*, 40–48. Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective, doi:https://doi.org/10.1016/j.sbi.2017.10.008.
29. Ahmed, M.C.; Papaleo, E.; Lindorff-Larsen, K. How well do force fields capture the strength of salt bridges in proteins? *PeerJ* **2018**, *6*, e4967. doi:10.7717/peerj.4967.
30. Piana, S.; Lindorff-Larsen, K.; Shaw, D.E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal* **2011**, *100*, L47–L49. doi:10.1016/j.bpj.2011.03.051.
31. Rieloff, E.; Skepö, M. Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison. *Int. J. Mol. Sci.* **2021**. accepted.
32. Cleveland, D.W.; Hwo, S.Y.; Kirschner, M.W. Purification of tau, a microtubule-associated protein that induces assembly of microtubules from purified tubulin. *Journal of Molecular Biology* **1977**, *116*, 207–225. doi:https://doi.org/10.1016/0022-2836(77)90213-3.
33. Buée, L.; Bussière, T.; Buée-Scherrer, V.; Delacourte, A.; Hof, P.R. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders11These authors contributed equally to this work. *Brain Research Reviews* **2000**, *33*, 95–130. doi:https://doi.org/10.1016/S0165-0173(00)00019-9.
34. Gong, C.X.; Iqbal, K. Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease. *Curr. Med. Chem.* **2008**, *15*, 2321–2328. doi:https://doi.org/10.2174/092986708785909111.
35. Hay, D.; Smith, D.; Schluckebier, S.; Moreno, E. Basic Biological Sciences Relationship between Concentration of Human Salivary Statherin and Inhibition of Calcium Phosphate Precipitation in Stimulated Human Parotid Saliva. *J. Dent. Res.* **1984**, *63*, 857–863.
36. Moreno, E.; Zahradnik, R. Demineralization and Remineralization of Dental Enamel. *JJ. Dent. Res.* **1979**, *58*, 896–903.
37. Raj, P.A.; Johnsson, M.; Levine, M.J.; Nancollas, G.H. Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization. *J. Biol. Chem.* **1992**, *267*, 5968–76.
38. Holt, C.; Timmins, P.A.; Errington, N.; Leaver, J. A core-shell model of calcium phosphate nanoclusters stabilized by β-casein phosphopeptides, derived from sedimentation equilibrium and small-angle X-ray and neutron-scattering measurements. *European Journal of Biochemistry* **1998**, *252*, 73–78, [https://febs.onlinelibrary.wiley.com/doi/pdf/10.1046/j.1432-1327.1998.2520073.x]. doi:https://doi.org/10.1046/j.1432-1327.1998.2520073.x.
39. Little, E.M.; Holt, C. An equilibrium thermodynamic model of the sequestration of calcium phosphate by casein phosphopeptides. *European Biophysics Journal* **2004**, *33*, 435–447. doi:10.1007/s00249-003-0376-x.
40. Ferraretto, A.; Gravaghi, C.; Fiorilli, A.; Tettamanti, G. Casein-derived bioactive phosphopeptides: role of phosphorylation and primary structure in promoting calcium uptake by HT-29 tumor cells. *FEBS Letters* **2003**, *551*, 92–98. doi:https://doi.org/10.1016/S0014-5793(03)00741-5.
41. Austin Elam, W.; Schrank, T.P.; Campagnolo, A.J.; Hilser, V.J. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Science* **2013**, *22*, 405–417, [https://onlinelibrary.wiley.com/doi/pdf/10.100 doi:https://doi.org/10.1002/pro.2217.
42. Andrew, C.D.; Warwicker, J.; Jones, G.R.; Doig, A.J. Effect of Phosphorylation on α-Helix Stability as a Function of Position. *Biochemistry* **2002**, *41*, 1897–1905, [https://doi.org/10.1021/bi0113216]. PMID: 11827536, doi:10.1021/bi0113216.
43. Errington, N.; Doig, A.J. A Phosphoserine-Lysine Salt Bridge within an α-Helical Peptide, the Strongest α-Helix Side-Chain Interaction Measured to Date. *Biochemistry* **2005**, *44*, 7553–7558, [https://doi.org/10.1021/bi050297j]. PMID: 15895998, doi:10.1021/bi050297j.
44. Gallivan, J.P.; Dougherty, D.A. Cation-πinteractions in structural biology. *Proceedings of the National Academy of Sciences* **1999**, *96*, 9459–9464, [https://www.pnas.org/content/96/17/9459.full.pdf]. doi:10.1073/pnas.96.17.9459.
45. Tayubi, I.; Sethumadhavan, R. Nature of cation-πinteractions and their role in structural stability of immunoglobulin proteins. *Biochemistry Moscow* **2010**, *75*, 912–918. doi:10.1134/S000629791007014X.
46. Das, R.K.; Pappu, R.V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences* **2013**, *110*, 13392–13397, [https://www.pnas.org/content/110/33/13392.full.pdf]. doi:10.1073/pnas.1304749110.
47. Kumar, P.; Chimenti, M.S.; Pemble, H.; Schönichen, A.; Thompson, O.; Jacobson, M.P.; Wittmann, T. Multisite Phosphorylation Disrupts Arginine-Glutamate Salt Bridge Networks Required for Binding of Cytoplasmic Linker-associated Protein 2 (CLASP2) to End-binding Protein 1 (EB1) *. *Journal of Biological Chemistry* **2012**, *287*, 17050–17064. doi:10.1074/jbc.M111.316661.
48. Berendsen, H.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56. doi:https://doi.org/10.1016/0010-4655(95)00042-E.

49. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447. PMID: 26620784, doi:10.1021/ct700301q.
50. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M.R.; Smith, J.C.; Kasson, P.M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854. doi:10.1093/bioinformatics/btt055.
51. Páll, S.; Abraham, M.J.; Kutzner, C.; Hess, B.; Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. Solving Software Challenges for Exascale; Markidis, S.; Laure, E., Eds.; Springer International Publishing: Cham, 2015; pp. 3–27.
52. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25. doi:https://doi.org/10.1016/j.softx.2015.06.001.
53. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958. doi:10.1002/prot.22711.
54. Piana, S.; Donchev, A.G.; Robustelli, P.; Shaw, D.E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123. PMID: 25764013, doi:10.1021/jp508971m.
55. Homeyer, N.; Horn, A.H.C.; Lanig, H.; Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281–289. doi:10.1007/s00894-005-0028-4.
56. Steinbrecher, T.; Latzer, J.; Case, D.A. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory Comput.* **2012**, *8*, 4405–4412. PMID: 23264757, doi:10.1021/ct300613v.
57. Hanwell, M.D.; Curtis, D.E.; Lonie, D.C.; Vandermeersch, T.; Zurek, E.; Hutchison, G.R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics* **2012**, *4*, 17. doi:10.1186/1758-2946-4-17.
58. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.2r1. The PyMOL Molecular Graphics System, version 1.2r1, 2009.
59. Berendsen, H.; Van Gunsteren, W., Practical algorithms for dynamic simulations. In *Molecular-Dynamics Simulations of Statistical-Mechanical Systems*; 1986; pp. 43–65.
60. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. doi:10.1063/1.464397.
61. Hess, B.; Bekker, H.; Berendsen, H.J.C.; Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
62. Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101. doi:10.1063/1.2408420.
63. Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190. doi:10.1063/1.328693.
64. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S.J.; Brett, M.; Wilson, J.; Millman, K.J.; Mayorov, N.; Nelson, A.R.J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.J.; Polat, İ.; Feng, Y.; Moore, E.W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E.A.; Harris, C.R.; Archibald, A.M.; Ribeiro, A.H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272. doi:10.1038/s41592-019-0686-2.
65. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637. doi:10.1002/bip.360221211.
66. Mansiaux, Y.; Joseph, A.P.; Gelly, J.C.; de Brevern, A.G. Assignment of PolyProline II Conformation and Analysis of Sequence – Structure Relationship. *PLOS ONE* **2011**, *6*, 1–15. doi:10.1371/journal.pone.0018401.
67. Chebrek, R.; Leonard, S.; de Brevern, A.G.; Gelly, J.C. PolyprOnline: polyproline helix II and secondary structure assignment database. *Database* **2014**, *2014*, [https://academic.oup.com/database/article-pdf/doi/10.1093/database/bau102/8248386/bau102.pdf]. bau102, doi:10.1093/database/bau102.
68. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. doi:10.1016/j.bpj.2015.08.015.
69. Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, M.; Ogasawara, H.; Näslund, L.Å.; Hirsch, T.K.; Ojamäe, L.; Glatzel, P.; Pettersson, L.G.M.; Nilsson, A. The Structure of the First Coordination Shell in Liquid Water. *Science* **2004**, *304*, 995–999, [https://science.sciencemag.org/content/304/5673/995.full.pdf]. doi:10.1126/science.1096205.
70. Campos, S.R.R.; Baptista, A.M. Conformational Analysis in a Multidimensional Energy Landscape: Study of an Arginylglutamate Repeat. *J. Phys. Chem. B* **2009**, *113*, 15989–16001. PMID: 19778072, doi:10.1021/jp902991u.
71. Henriques, J.; Cragnell, C.; Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 3420–3431, [https://doi.org/10.1021/ct501178z]. PMID: 26575776, doi:10.1021/ct501178z.
72. Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
73. Stone, J. An Efficient Library for Parallel Ray Tracing and Animation, 1998. Master thesis.

74. Frishman, D.; Argos, P. Knowledge-based secondary structure assignment. *Proteins* **1995**, *23*, 566–579.
75. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C. Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas; Loizides, F.; Schmidt, B., Eds. IOS Press, 2016, pp. 87–90.
76. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M.H.; Brett, M.; Haldane, A.; Fernández del Río, J.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T.E. Array programming with NumPy. *Nature* **2020**, *585*, 357—-362. doi:10.1038/s41586-020-2649-2.
77. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **2007**, *9*, 90–95. doi:10.1109/MCSE.2007.55.

Supplementary File S1 for Conformational effects of multisite phosphorylation of disordered peptides

E. Rieloff[1] and M. Skepö[1,2*]

[1] Division of Theoretical Chemistry, Lund University, Lund, Sweden
[2] LINXS - Lund Institute of Advanced Neutron and X-ray Science, Lund, Sweden

* marie.skepo@teokem.lu.se

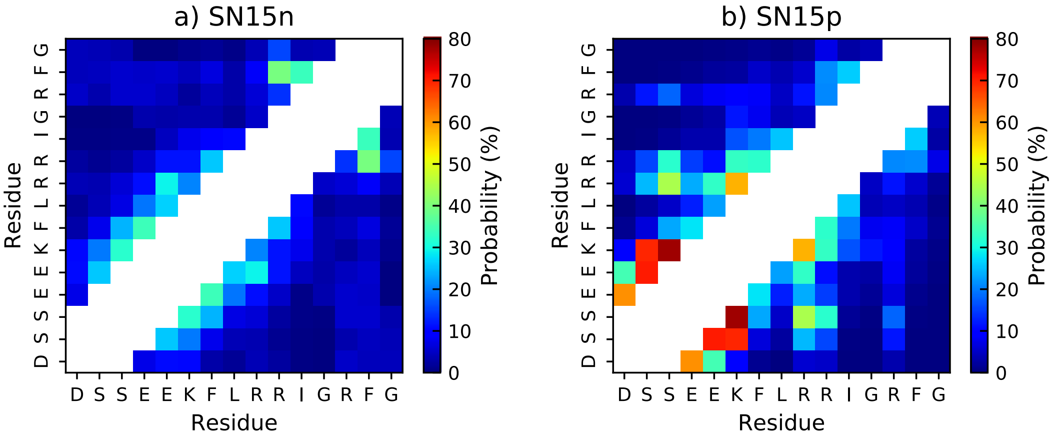


Figure S1: Contact map of a) non-phosphorylated and b) phosphorylated SN15. Probability of atoms in different residues being closer than 4 Å, with the two closest residues on each side as well as the residue itself excluded from analysis and therefore shown in white. The data are taken from ref. [1].

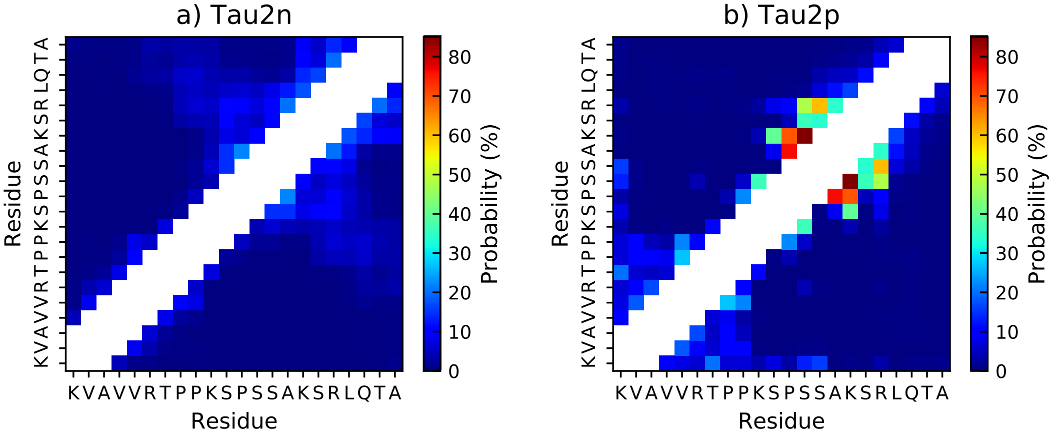


Figure S2: Contact map of a) non-phosphorylated and b) phosphorylated Tau2. Probability of atoms in different residues being closer than 4 Å, with the two closest residues on each side as well as the residue itself excluded from analysis and therefore shown in white.

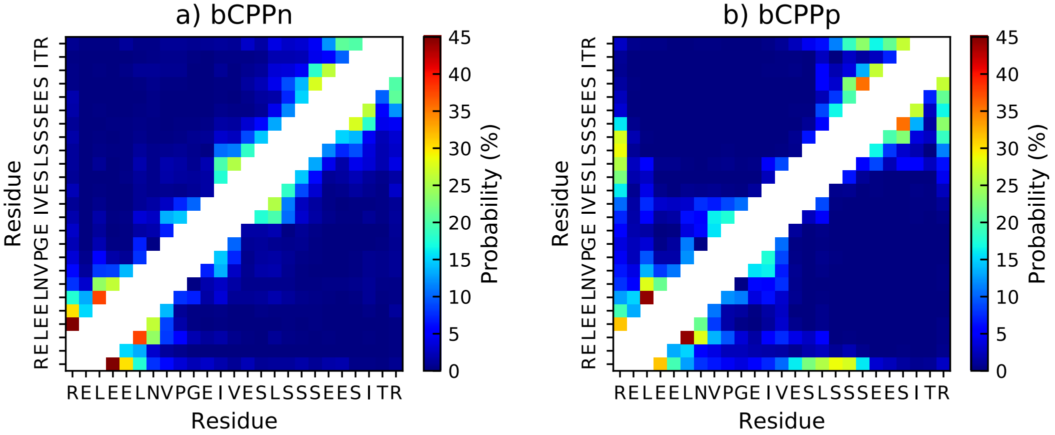


Figure S3: Contact map of a) non-phosphorylated and b) phosphorylated bCPP. Probability of atoms in different residues being closer than 4 Å, with the two closest residues on each side as well as the residue itself excluded from analysis and therefore shown in white.

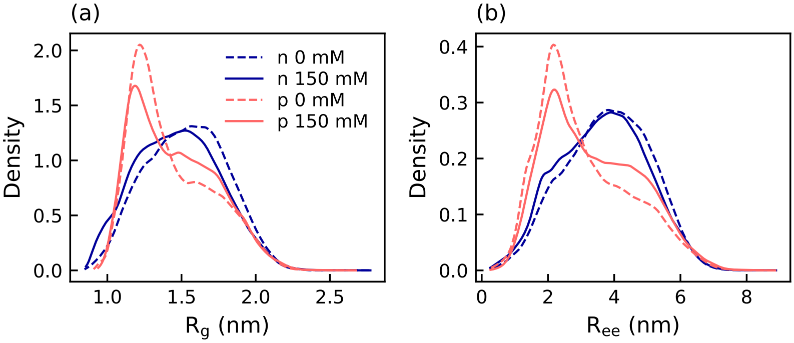


Figure S4: Distribution of a) radius of gyration and b) end-to-end distance of non-phosphorylated (n) and phosphorylated (p) bCPP simulated with 0 or 150 mM NaCl.

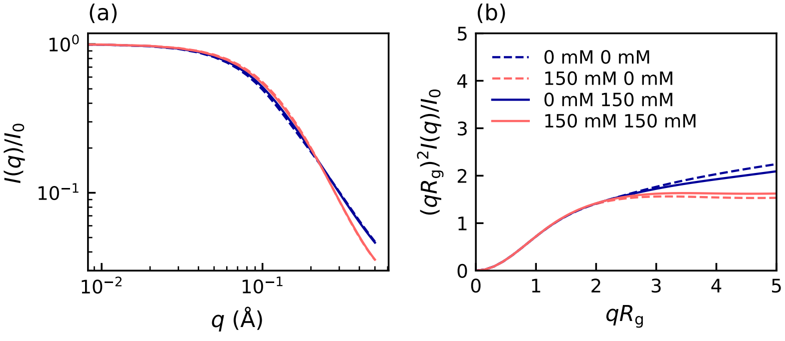


Figure S5: a) Calculated form factor and b) dimensionless Kratky plot of non-phosphorylated (n) and phosphorylated (p) bCPP simulated with 0 or 150 mM NaCl.

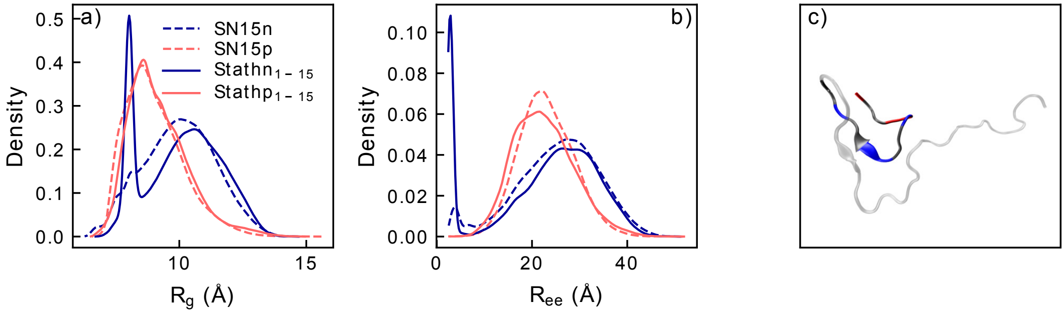


Figure S6: a) Radius of gyration and b) end-to-end distance distributions of non-phosphorylated (n) and phosphorylated (p) Stath$_{1\text{-}15}$ and SN15. c) Snapshot of the type of conformation giving rise to the sharp peak in the Stathn$_{1\text{-}15}$ distributions, where residue 16-43 is traced in light gray. The SN15 data are taken from ref. [1].

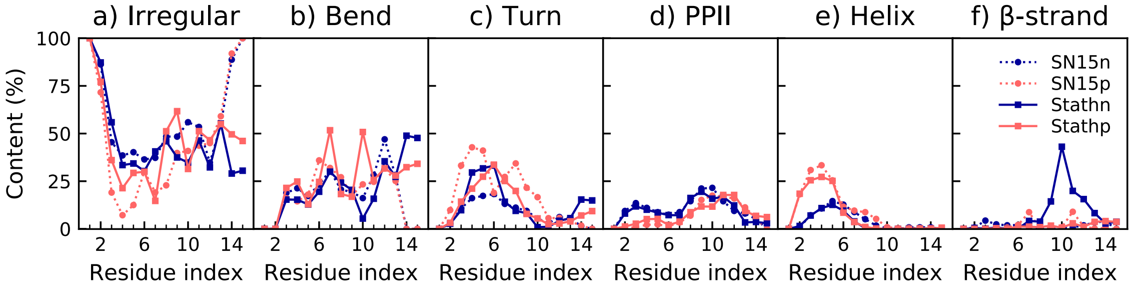


Figure S7: Secondary structure content along the sequence in non-phosphorylated (n) and phosphorylated (p) Stath$_{1\text{-}15}$ and SN15. Helix includes α-helix and 3_{10}-helix. β-strand also includes β-bridge. The SN15 data are taken from ref. [1].

References

[1] Ellen Rieloff and Marie Skepö. Phosphorylation of a disordered peptide—structural effects and force field inconsistencies. *Journal of Chemical Theory and Computation*, 16(3):1924–1935, 2020. PMID: 32050065.

Supplementary File S2 for Conformational effects of multisite phosphorylation of disordered peptides

E. Rieloff[1] and M. Skepö[1,2*]

[1] Division of Theoretical Chemistry, Lund University, Lund, Sweden
[2] LINXS - Lund Institute of Advanced Neutron and X-ray Science, Lund, Sweden

* marie.skepo@teokem.lu.se

The time evolution, density distribution, autocorrelation function, and error estimate from block averaging of end-to-end distance and radius of gyration, as well as the energy landscapes constructed from principal component analysis have all been used to assess the convergence and sampling quality of the simulations in this work. For SN15n and SN15p we refer to the supporting information of ref. [1] and for Tau2p, bCPPp, and Stathp we refer to the supplementary material to ref. [2]. The remaining peptides are presented in the following order:

- Tau1n: Figure A1–A3
- Tau1p: Figure A4–A6
- Tau2n: Figure A7–A9
- bCPPn: Figure A10–A12
- bCPPn 150 mM: Figure A13–A14
- Stathn: Figure A15–A17

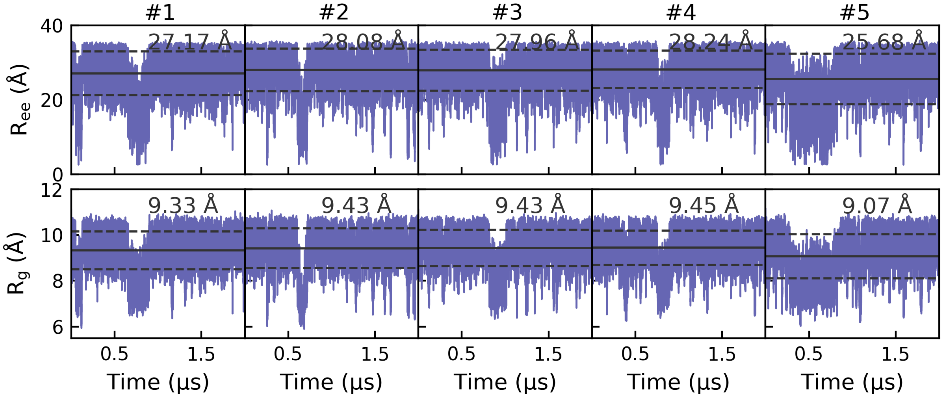


Figure A1: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Tau1n. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

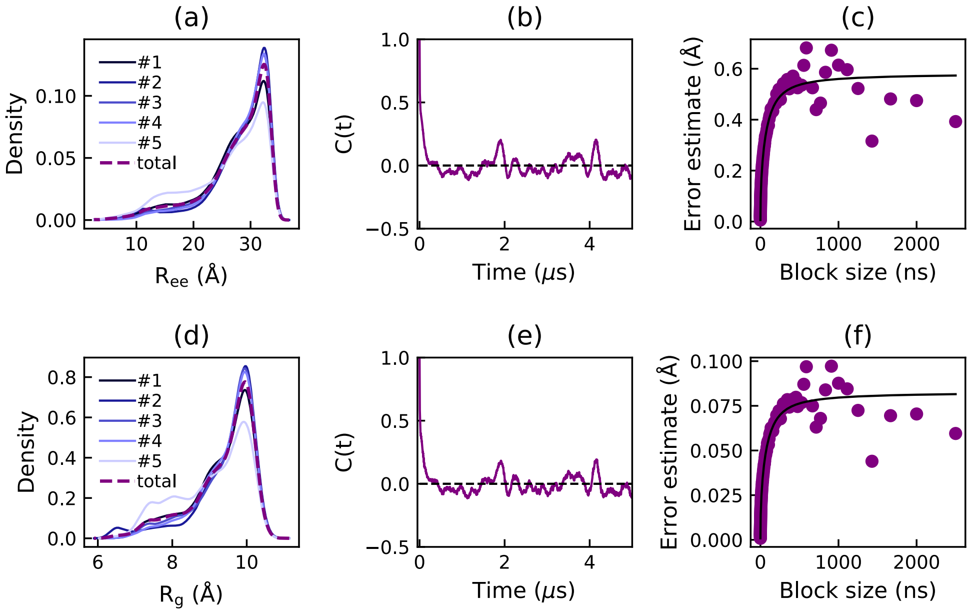


Figure A2: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of Tau1n, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

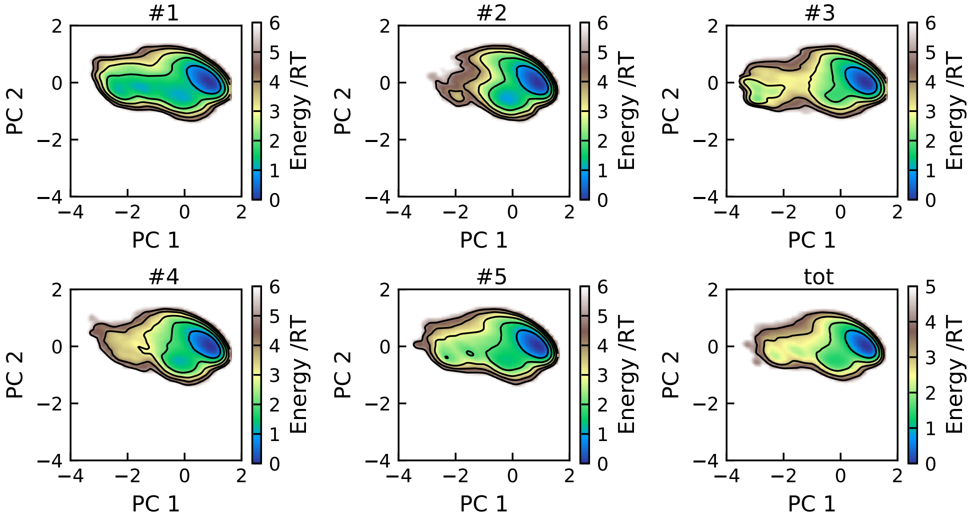


Figure A3: Energy landscapes for the five replicates and the concatenated trajectory of Tau1n, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

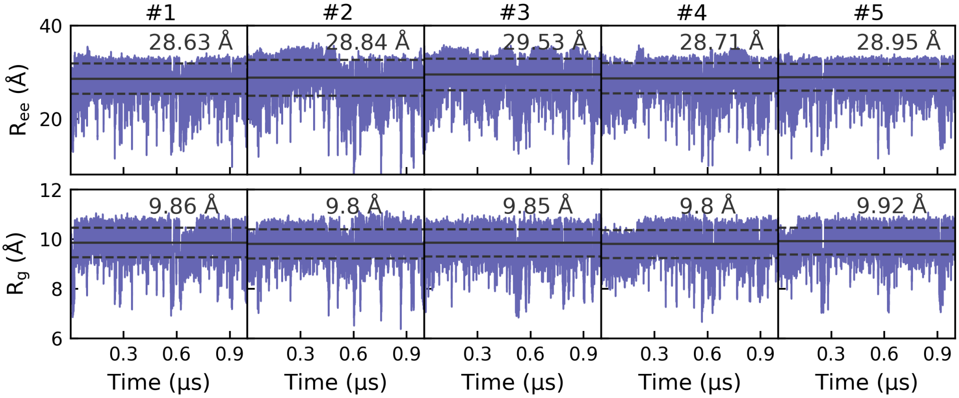


Figure A4: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Tau1p. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

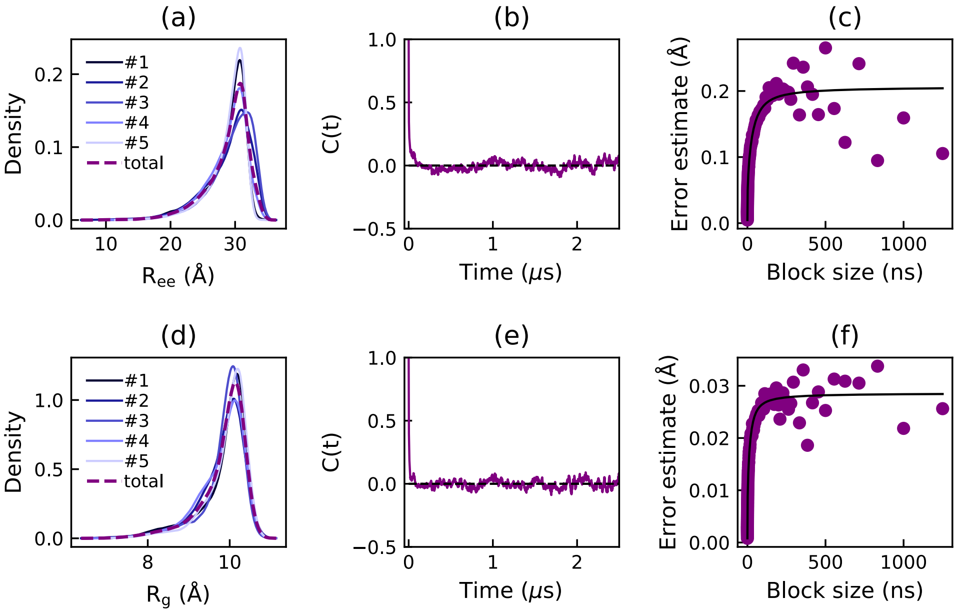


Figure A5: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of Tau1p, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

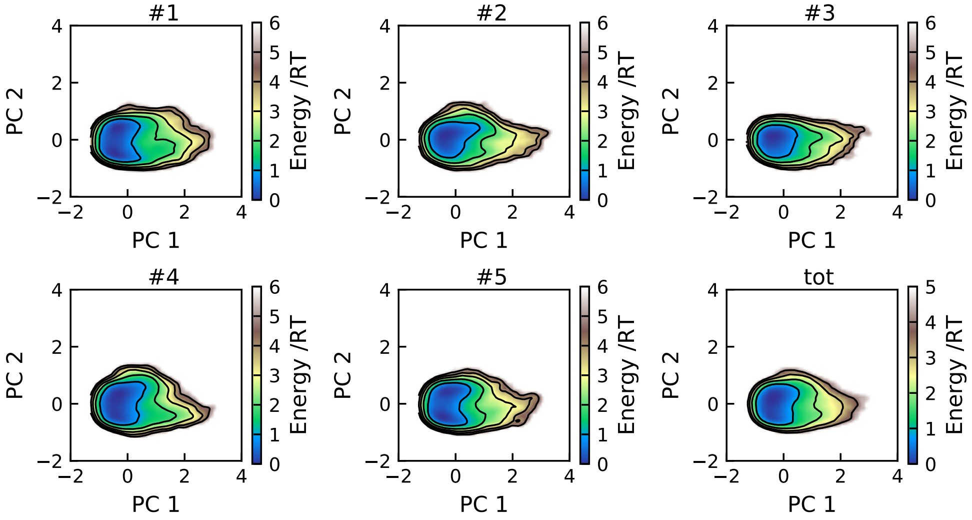


Figure A6: Energy landscapes for the five replicates and the concatenated trajectory of Tau1p, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

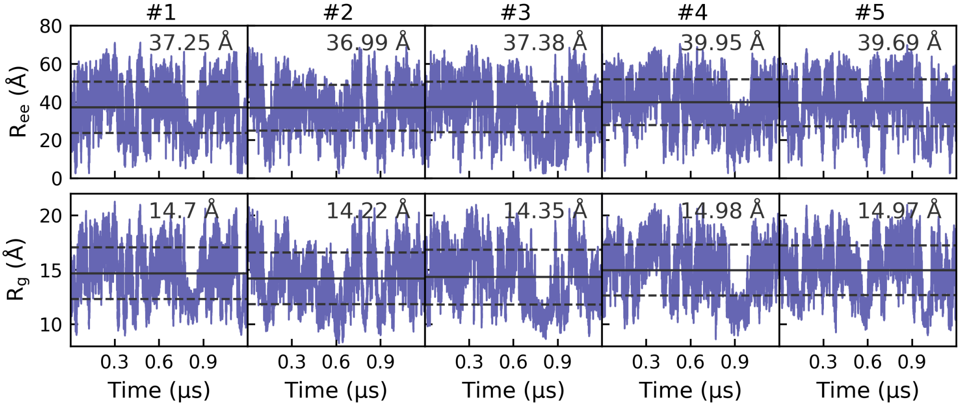


Figure A7: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Tau2n. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

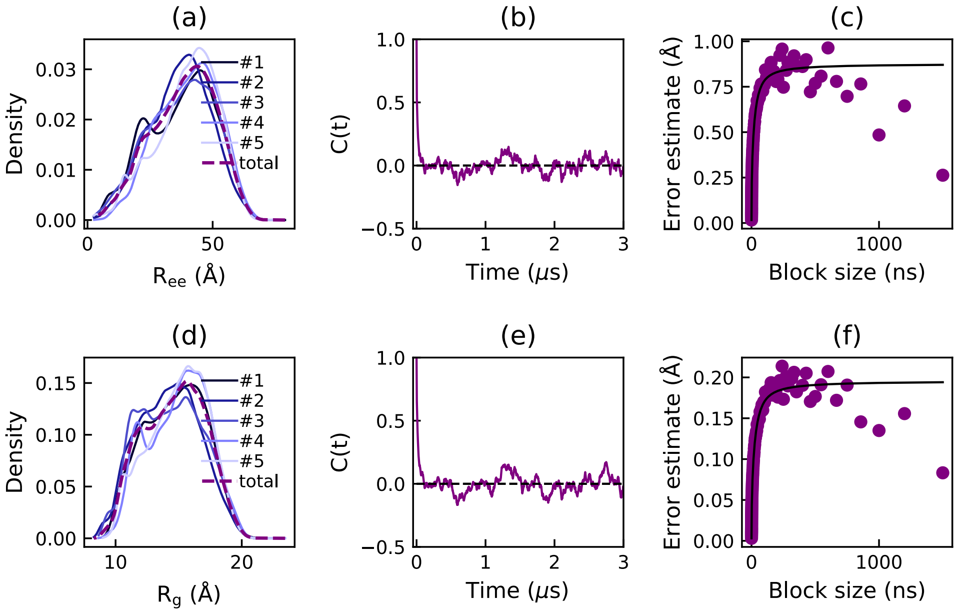


Figure A8: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of Tau2n, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

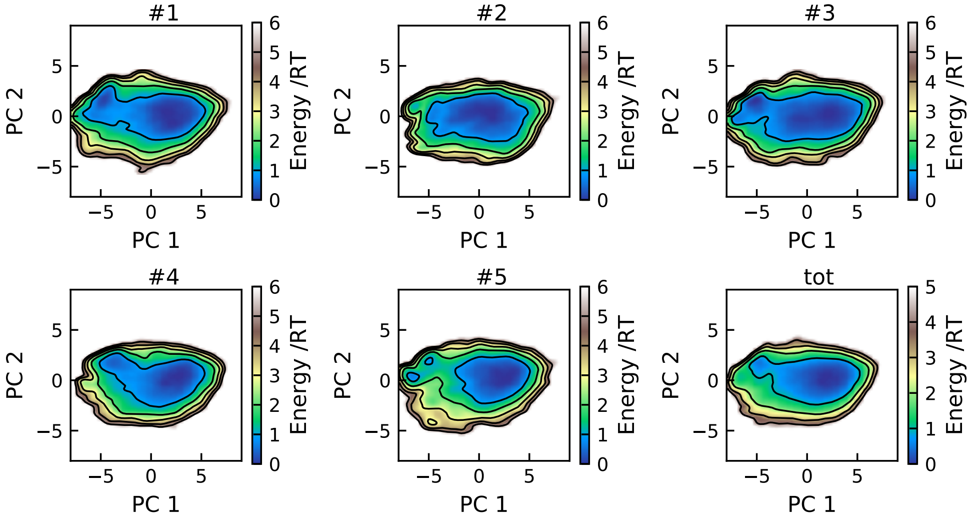


Figure A9: Energy landscapes for the five replicates and the concatenated trajectory of Tau2n, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

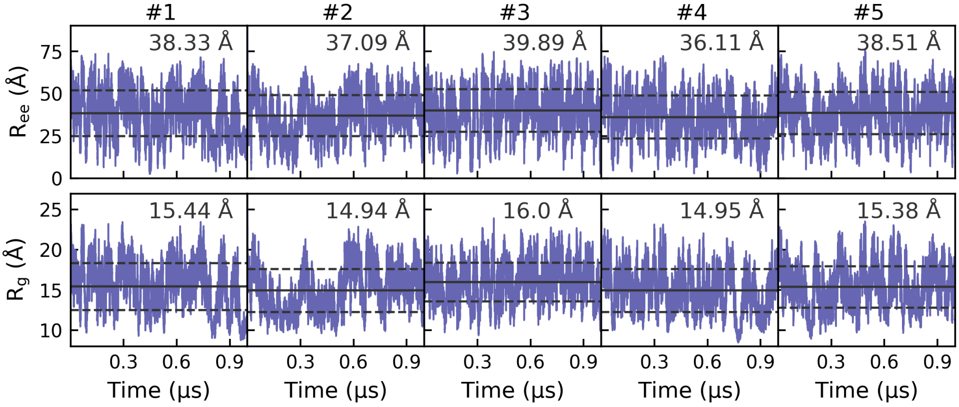


Figure A10: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_{g}) for the five replicates in the simulation of bCPPn. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

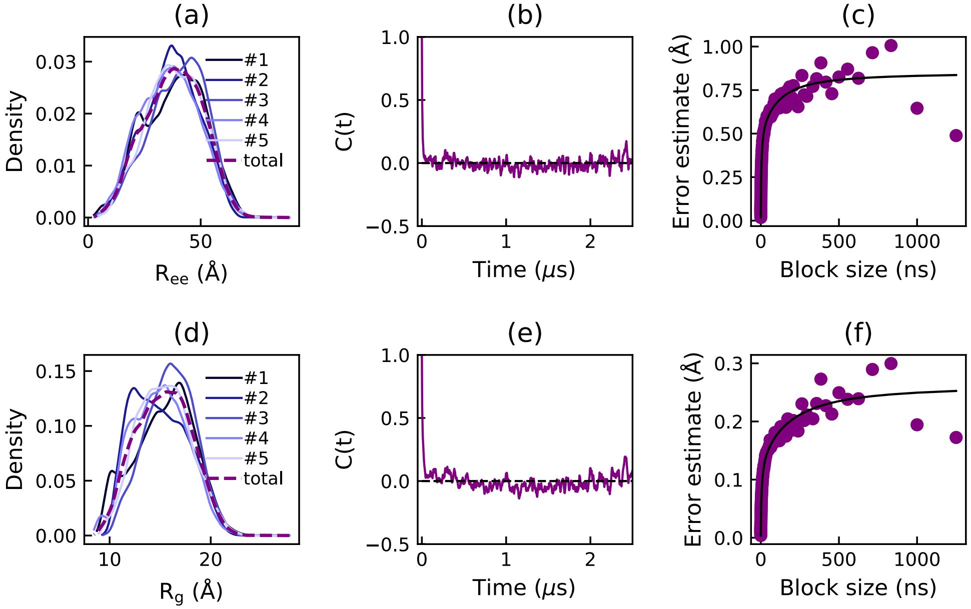


Figure A11: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of bCPPn, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

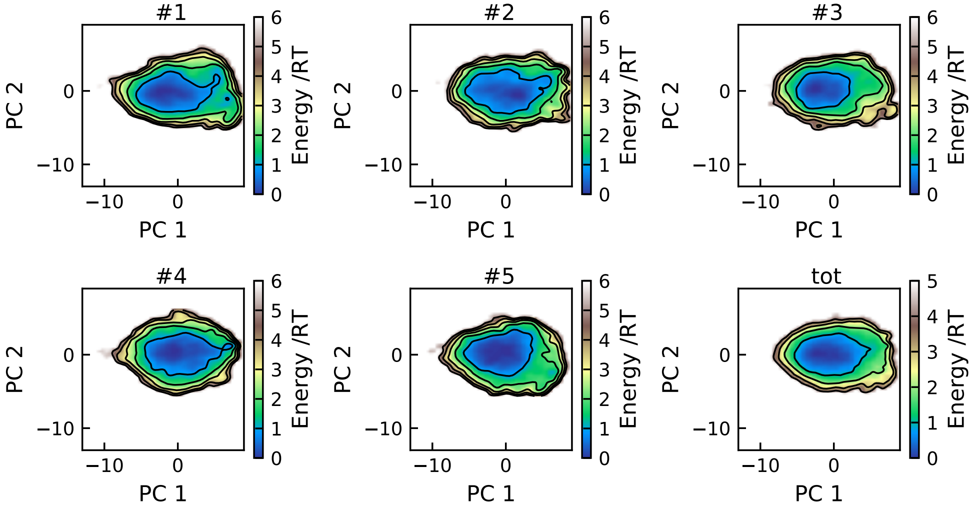


Figure A12: Energy landscapes for the five replicates and the concatenated trajectory of bCPPn, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

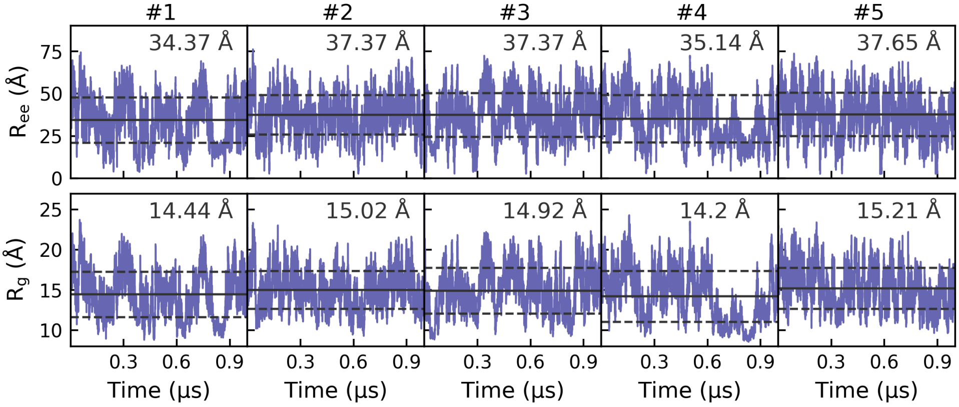


Figure A13: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of bCPPn with 150 mM NaCl. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

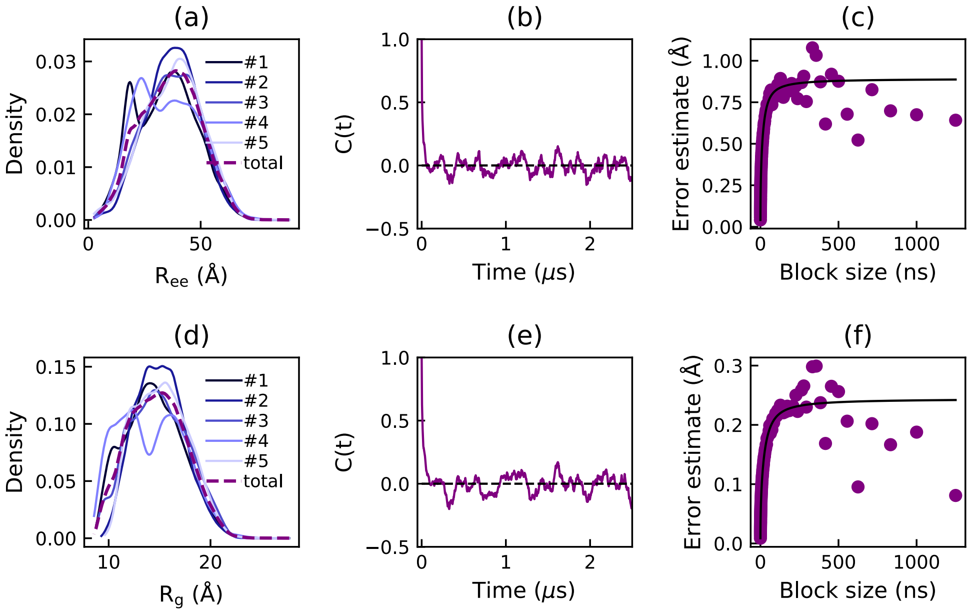


Figure A14: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of bCPPn with 150 mM NaCl, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

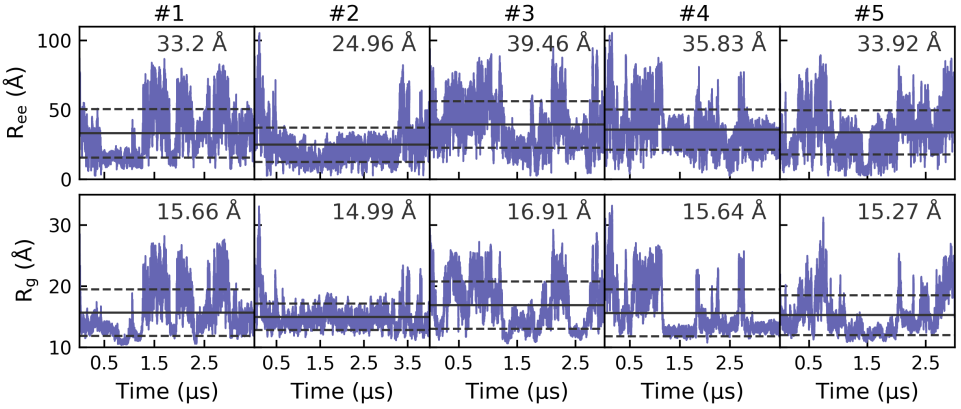


Figure A15: Time evolution of the end-to-end distance (R_{ee}) and the radius of gyration (R_g) for the five replicates in the simulation of Stathn. The horizontal solid line represents the average in each replicate, with the dashed lines showing the standard deviation.

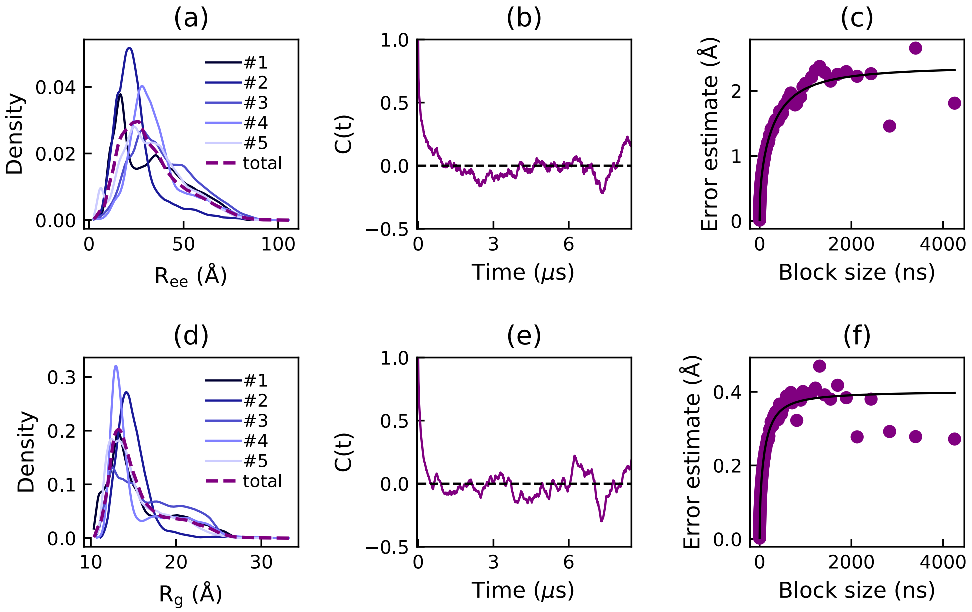


Figure A16: Density estimates of the end-to-end distance (R_{ee}) (a) and the radius of gyration (R_g) (d) for the five replicates and the concatenated simulation of Stathn, obtained from a Gaussian kernel estimator. Autocorrelation function (C(t)) and error estimate from block averaging of the end-to-end distance (b,c) and the radius of gyration (e,f) for the concatenated simulation.

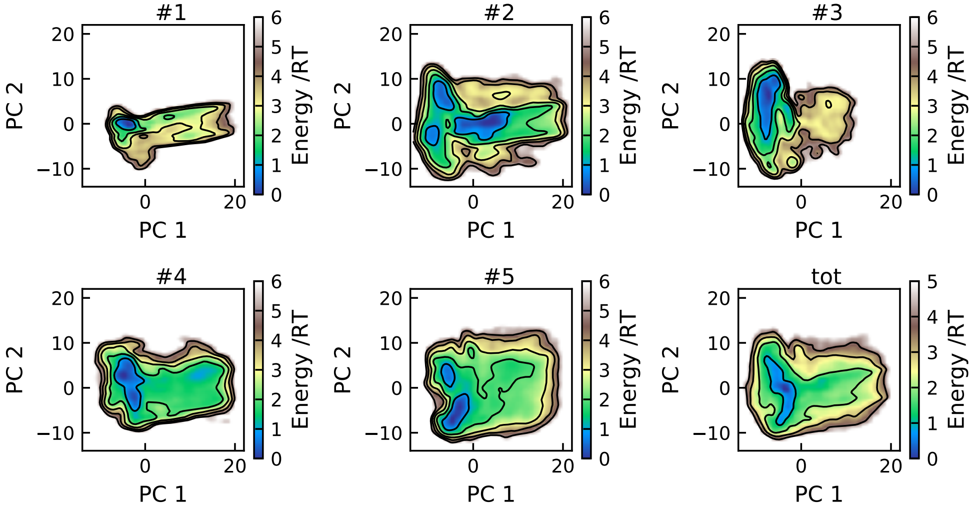


Figure A17: Energy landscapes for the five replicates and the concatenated trajectory of Stathn, using the first two principal components. All plots have been constructed using the same basis set and are therefore directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$.

References

[1] Ellen Rieloff and Marie Skepö. Phosphorylation of a disordered peptide—structural effects and force field inconsistencies. *J. Chem. Theory Comput.*, 16(3):1924–1935, 2020. PMID: 32050065.

[2] Ellen Rieloff and Marie Skepö. Molecular dynamics simulations of phosphorylated intrinsically disordered proteins: A force field comparison. *Int. J. Mol. Sci.*, 2021. accepted.

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

In this thesis, computational and experimental methods are applied to study the conformational ensembles of intrinsically disordered proteins. The main goals have been to investigate the relation between sequence and structure, focusing on the impact of phosphorylation, and to investigate different models applicable for studying intrinsically disordered proteins.

LUND
UNIVERSITY

ISBN: 978-91-7422-828-1

Department of Chemistry
Faculty of Science
Lund University

9 789174 228281