Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

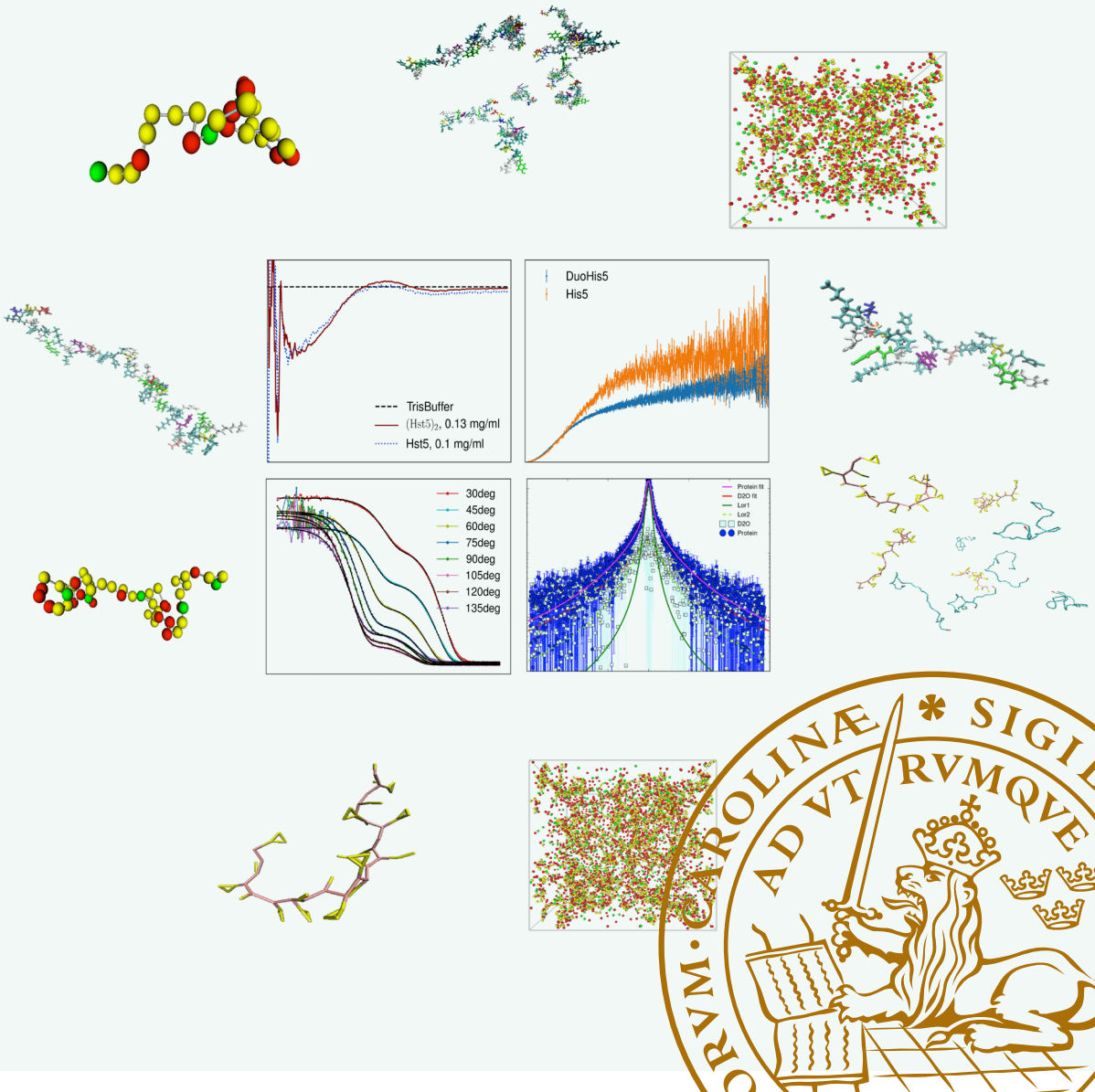
Fagerberg, Eric

2022

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

Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

ERIC FAGERBERG | DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY



Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

by Eric Fagerberg

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Prof. Marie Skepö and Dr. Tilo Seydel
Faculty opponent: Dr. Alex Holehouse, Washington University School of Medicine, Dept. of Biochemistry & Molecular Biophysics

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in the KC:A lecture hall at the Department of Chemistry on Friday, the 7th of October 2022 at 13:00.

Organization
LUND UNIVERSITY

Department of Chemistry
Box 124
SE–221 00 LUND
Sweden

Author(s)
Eric Fagerberg

Document name
DOCTORAL DISSERTATION

Date of disputation
2022-10-07

Sponsoring organization
Vinnova, Crafoord Foundation, Royal Physiographic Society of Lund

Title and subtitle

Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

Abstract

Intrinsically disordered proteins are distinguished by a lack of distinct three-dimensional structure, existing instead as an ensemble of heterogenous structures. In this research, the effect of crowding on these proteins is investigated using a combined approach of experiment and computer simulation, mainly using coarse-grained simulation models to make simulation computationally feasible at the high concentration conditions crowding is displayed.

Firstly, the saliva protein Histatin 5 (Hst5) is studied with SAXS, where a selection of coarse-grained models were evaluated using the SAXS data. It was determined that no model could provide adequate simulation-experiment agreement, but a best-performing model could be established. This model predicted moderate change in structure with crowding in the case of Histatin 5.

It was postulated the moderate effect of crowding on Histatin 5 was due to its short sequence-length. Thus, the dimer of Hst5 was formed and subjected to investigation by SAXS and computer simulation for crowding effects. The dimer was more challenging to model with a coarse-grained model, and circular dichroism data suggested secondary structures to be present, which a coarse-grained model cannot capture. Atomistic modelling followed, which however did not perform better than the coarse-grained models, showing the importance of further developing these models to represent intrinsically disordered proteins.

Atomistic modelling was also performed at high concentrations of Hst5 5, combined with quasi-elastic neutron spectroscopy to elucidate diffusion behaviour at crowded conditions. Diffusion decreased with increasing protein concentration, with temperature effects following Stokes-Einstein behaviour and increses in salt content to decrease diffusion. Depending on assumptions on the relation between effective- and translational-diffusion, the atomistic model displayed semi-quantitative agreement with experiment.

Using neutral polymeric crowders rather than self-crowding showed no impact on structure, as investigated by SAXS. Using DLS did as well not reveal any crowding impact, with the exception of Ficoll®, where Hst5 seemed to modulate Ficoll® self-crowding behaviour in terms of diffusion, decreasing the self-crowding effect. Several coarse-grained models showed similar non-existant effects on structure by crowding, with small deviations from experiment.

Benchmarking three coarse-grained models indicate higher degree of finegraining and additional parameters does necessarily follow the intuitive notion of increasing performance, with the most advanced not having as good performance as the two simpler models in terms of predicting radius of gyration.

Key words
simulation, IDP, crowding

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-7422-898-4 (print)
978-91-7422-899-1 (pdf)

Recipient's notes

Number of pages
258

Price

Security classification

Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

by Eric Fagerberg

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 summarizes 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).

Cover illustration front: Snapshots from the different simulations performed, using different kinds of models. The middle four pictures are from the different experimental techniques used (from top left: circular dichroism, small angle X-ray scattering, dynamic light scattering, and quasi-elastic neutron scattering).

Funding information: The thesis work was financially supported Vinnova, Crafoord Foundation, Royal Physiographic Society of Lund and the foundation in memory of Per Westling

Parts of this thesis has been published before in:
Self-crowding of unstructured proteins studied by small angle X-ray scattering and coarse-grained simulations (2020)

Faculty of Science, Department of Chemistry

ISBN: 978-91-7422-898-4 (print)
ISBN: 978-91-7422-899-1 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2022

Preface

The text in your hands is a product of approximately four years worth of work. I would like to give some reasons for why I wanted to spend that time on this text.

Firstly, there is a bit of a ”game” in making predictions and see how they compare with experiment. The game is made interesting here by not always being a binary question of ”was prediction close enough to the experimental number”, but one can look at different aspects, such as predictions of varying systems or varying molecular properties.

Secondly, I have a fascination of the underlying theory of the models used here, the framework of statistical thermodynamics. With a few assumptions, there is a claim of being able to predict ALL (thermodynamic) properties of any given system. Methods starting off in this framework thus have the opportunity of being universal, being the ultimate source of knowledge. Certainly, it is a very pretentious view of things, and probably impossible to practically achieve (as a joke, I tell people that experimental work will be obsolete in 50 years, unfortunately people don't laugh at this statement), but it is absolutely fascinating.

Lastly, I like the craft. Using computers, scripting/programming, data analysis, and a bit of lab work from time to time is enjoyable, which should not be underestimated as a driving force.

Contents

List of publications

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

I **Evaluating models of varying complexity of crowded intrinsically disordered protein solutions against SAXS**
E. Fagerberg, S. Lenton, M. Skepö
Journal of Chemical Theory and Computation, 2019, 5(12), pp. 6968–-6983

II **The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions**
E. Fagerberg, L.K. Månsson, S. Lenton, M. Skepö
The Journal of Physical Chemistry B, 2020, 124(52), pp. 11843–11853

III **Self-Diffusive Properties of the Intrinsically Disordered Protein Histatin 5 and the Impact of Crowding Thereon: A Combined Neutron Spectroscopy and Molecular Dynamics Simulation Study**
E. Fagerberg, S. Lenton, T. Nylander, T. Seydel, M. Skepö
The Journal of Physical Chemistry B, 2022, 126(4), pp. 789-801

IV **The crowding effect using neutral crowders on Histatin 5. Computer simulations in combination with X-ray and dynamic light scattering**
E. Fagerberg, P. Holmqvist, S. Lenton, P. Pernot, M. Skepö
In preparation

V **Comparative performance of coarse-grained IDP models at different resolutions**
E. Fagerberg, M. Skepö
In preparation

Publications not included in this thesis:

PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins

Tamas Lazar, Elizabeth Martínez-Pérez, Federica Quaglia, András Hatos, Lucía B Chemes, Javier A Iserte, Nicolás A Méndez, Nicolás A Garrone, Tadeo E Saldaño, Julia Marchetti, Ana Julia Velez Rueda, Pau Bernadó, Martin Blackledge, Tiago N Cordeiro, **Eric Fagerberg**, Julie D Forman-Kay, Maria S Fornasari, Toby J Gibson, Gregory-Neal W Gomes, Claudiu C Gradinaru, Teresa Head-Gordon, Malene Ringkjøbing Jensen, Edward A Lemke, Sonia Longhi, Cristina Marino-Buslje, Giovanni Minervini, Tanja Mittag, Alexander Miguel Monzon, Rohit V Pappu, Gustavo Parisi, Sylvie Ricard-Blum, Kiersten M Ruff, Edoardo Salladini, Marie Skepö, Dmitri Svergun, Sylvain D Vallet, Mihaly Varadi, Peter Tompa, Silvio C E Tosatto, Damiano Piovesan

Nucleic Acids Research, Vol. 49, Issue D1, 2021, pp. D404-D411

Acknowledgements

I would like to thank my supervisor Marie Skepö, for taking me in as a PhD student and advicing me throughout this time period. Additionally, help from my co-supervisor Tilo Seydel has been much appreciated. As well, thanks are given to past and present members of our group that I had the pleasure to become acquainted with - Carolina, Kristin, Linda, Maria, Stephanie, Ellen, Sandeep, Jenny, Amanda, Mona - for both fruitful discussions and helpful tips. Special thanks to Samuel Lenton, who taught me about SAXS and for putting up with me asking thousands of questions about everything.

Also, I really appreciate everyone at the division for making the workplace a very pleasant place to be at.

Last but not least, I would like to thank friends and family for everything else.

Popular summary in English

Proteins are more than something to eat to gain muscles, they are necessary for the body to function at all. For proteins to have a function it was long believed that the protein needed a well-specified 3D-structure. This view has changed after so-called intrinsically disordered proteins (IDPs) were confirmed to have important biological functions. Estimates also show that they are not a fringe-class of proteins, but encompass almost 30 % of proteins in eukaryotic organisms.

Distinguishing IDPs from other proteins is the very fact that they lack a specific 3D-structure. Instead they exist as ensembles of heterogenous structures. However, this makes them harder to study experimentally, why computer simulations may contribute to the understanding of these proteins.

In many studies, proteins are investigated at low protein concentrations. But in biological contexts, the concentration of macromolecules (such as proteins, fats, carbohydrates, etc) is high. The high concentration may affect proteins to assume other structures than those observed at lower concentrations, thus, studies aiming to investigate biological function should consider this effect.

In this Thesis, a particular protein found in the saliva, Histatin 5, has been studied at high concentrations using scattering methods and computer simulations. Having high concentration is an impediment for simulations, as it is more computationally demanding. To get around this problem, so-called coarse-grained models has been utilized, which simplifies structures by putting several particles in groups, which then is treated as one big particle.

Several coarse-grained models were considered together with experimental data at high Histatin 5 protein concentration, and while no model was perfect, the best performing model indicated that concentration effects could be found up to medium-high concentrations. At very high concentration, experiments indicated that Histatin 5 to some extent lumped together to form aggregates, which simulations did not indicate. An hypothesis for the experimental behaviour was that Histatin 5 was too small protein for any dramatic effects to be observed, why a new protein was constructed by putting together two Histatin 5 proteins, attaching the C-terminal of the first Histatin 5 protein with the N-terminal of the second Histatin 5 protein. The same experiment was repeated, together with the best-performing model from before, and it was found that for the repeat-protein, the model was worse performing, even at lower protein concentrations. Some experiments pointed to properties in the repeat-protein that could not be modelled with the previous model, why a more powerful but more computationally expensive, non-coarse-grained (atomistic) model, able to model additional properties, was applied, but it was found that even this model had difficulty explaining experimental data. This shows the need to develop models further to work better with this important class of proteins. Experimental behaviour of

the repeat-protein was as with Histatin 5, with the exception that the concentration where aggregation would start to be observed was lower than for Histatin 5.

In biological contexts, proteins are not surrounded by copies of themselves, but exist in a heterogenous environment. One should therefore also study proteins together with other sorts of molecules, that contribute to a crowded environment. Such a molecule, that are intended to make an environment crowded at increasing concentration, is aptly called a crowder. Here, Histatin 5 was studied together, each by themselves, together with four other molecules, in various size. For three of these, no effect on Histatin 5 was found at all, while the fourth, largest crowder had, according to experiments where structure was probed, no effect, while other experiments probing diffusion showed a difference when the concentration of crowder was fairly large. Simulations were able to confirm the lack of effect by the crowders concerning structural properties.

How diffusion was affected by high concentrations of Histatin 5 alone was also studied. It was found that diffusion slowed down, which possibly is explained by Histatin 5 lumping together at high concentrations. The effect of temperature and salt at high protein concentration was also investigated, where the temperature effect was found to be trivial (high temperature yielding faster diffusion), while increasing salt content decreased the diffusion rate. Speculatively, the salt effect was explained by Histatin 5 having a slightly different structure at low salt content, or that salt induces clustering of Histatin 5.

The examination of diffusion as a function of protein concentration also included simulations at the atomistic level. To compare experiments with the simulations, assumptions was needed. These were primarily chosen based on previous studies of another protein. With such assumptions, the simulation found too slow diffusion rates as compared with experiment. Other assumptions, based on approximations about the geometry of Histatin 5 would yield more favourable values compared with experiment, but studies on real proteins should be deemed as more realistic. The trends of diffusion as a function of protein concentration were similar between experiments and simulation, why the simulation model should be regarded as semi-quantitative.

Finally, a comparison between three different coarse grained models was performed, where it was found that the most advanced model did not perform as well as the simpler models in terms of predicting the overall dimensions of a set of intrinsically disordered proteins, indicating that simpler models are competitive.

Populärvetenskaplig sammanfattning på svenska

Proteiner är mer än något man behöver äta för att få stora muskler, de är absolut nödvändiga för att kroppen ska fungera överhuvudtaget. För att proteiner skulle ha en funktion ansågs det länge vara nödvändigt att proteinet hade en väl specifierad 3D-struktur. Denna syn har ändrats efter att så kallade intrinsikalt oordnade proteiner (”Intrinsically Disordered Protein”, förkortat IDP på engelska) bekräftats ha viktiga biologiska funktioner. Uppskattningar visar också att de inte är en liten grupp specialfall av proteiner, utan nästan 30 % av proteinerna i eukaryota organismer är IDP:er.

Det som utmärker IDP:er är just att de saknar en specifik 3D-struktur. Istället existerar dessa som en ensemble av heterogena strukturer. Det gör dem dock svårare att studera experimentellt, varför datorsimuleringar kan bidra till att förstå dessa proteiner.

I många studier studerar man proteiner i låga koncentrationer. Men i verkliga biologiska sammanhang är koncentrationen inte låg utan hög. Den höga koncentrationen kan få proteiner att anta andra strukturer än dem som observeras vid låga koncentrationer, så studier som syftar till att förstå biologisk funktion hos proteiner bör ta hänsyn till denna effekt.

I den här avhandlingen har speciellt ett protein som finns i saliven, Histatin 5, studerats vid höga koncentrationer med spridningstekniker och datorsimuleringar. Höga koncentrationer är i datorsimuleringar ett hinder, eftersom det är mer beräkningskrävande. För att komma runt detta har så kallade grovkorniga datormodeller använts, som förenklar strukturer genom att sätta flera partiklar i grupper, som sedan behandlas som en stor partikel.

Flera sådana grovkorniga modeller utvärderades tillsammans med experimentella data, och även om ingen modell var perfekt, visade den bästa modellen att för Histatin 5 finns ingen större effekt upp till medelhöga koncentrationer. Vid än högre koncentrationer visade experiment att Histatin 5 i viss mån började att klumpa ihop sig, vilket modellering inte visade. En hypotes för det experimentella beteende var att Histatin 5 var ett för litet protein för att mer dramatiska effekter skulle kunna ske, varför ett nytt protein där man satte ihop två Histatin 5 proteiner skapades och samma experiment upprepades, med den modell som i föregående fall presterat bäst.. Här var modellen sämre, även vid låga proteinkoncentrationer, och vissa experiment pekade på egenskaper hos det nya proteinet som inte kunde modelleras med den förenklade modellen. Därför utfördes simuleringar med en kraftfull, men beräkningskrävande modell som kan förklara fler egenskaper, men det visade sig att även denna hade problem med att förklara experimentella data. Detta visar på behovet att utveckla modeller vidare för att fungera bättre med denna viktiga klass av proteiner. Experimentellt var beteendet som tidigare, med undantaget att den koncentration där proteinet började klumpa sig var lägre.

I biologiska sammanhang är inte proteiner omgivna av kopior av sig själva, utan befinner

sig i en heterogen miljö. Man bör därför också studera proteiner tillsammans med andra sorters molekyler. Här studerades Histatin 5 var för sig ihop med fyra andra molekyler, i varierande storlekar. För tre av dessa fanns ingen effekt på Histatin 5 alls, medan den fjärde, största molekylen enligt ett experiment där struktur undersöks inte gav någon observerbar effekt, medan ett annat experiment där diffusion undersöks påvisade en skillnad när koncentrationen av molekylen blev större. Simuleringar kunde bekräfta bristen på effekt gällande strukturella egenskaper.

Hur diffusion påverkas av höga koncentrationer av Histatin 5 ensamt studerades också. Hög koncentration visades sakta ner diffusionstakten, vilket eventuellt kan förklaras att Histatin 5 klumpar ihop sig vid högre koncentrationer. Effekten av salt och temperatur vid höga koncentrationer undersöktes också, där temperatur visade sig påverka Histatin 5 trivialt (hög temperatur gav snabbare diffusion), medan salt minskade diffusionstakten, vilket spekulativt förklaras med att Histatin 5 har en något annorlunda struktur vid lägre salthalter eller att höga saltkoncentrationer får Histatin 5 att klumpa ihop sig.

För undersökningen av diffusion som funktion av koncentration användes simuleringar med hög detaljnivå. För att jämföra simuleringarna med experimenten behövdes dock några antaganden, vilka valdes utifrån tidigare studier av ett annat protein. Med det valda antagandet gav simulering en för långsam diffusion jämfört med experiment. Andra antaganden, som baserar sig på approximationer om geometrin hos Histatin 5 ger mer jämförbara värden med experiment, men det bör anses att studier på verkliga proteiner ger mer realistiska antaganden. Trender inom diffusion var däremot likartade gentemot experiment, varför simuleringsmodellen bör anses semi-kvantitativ.

Till sist testades flera grovkorniga modeller, där det visade sig att den mest avancerade modellen var den som presterade sämst, vilket visar på att enklare modeller är konkurrenskraftiga.

Assessing the structural and dynamical properties of concentrated solutions of the disordered proteins Histatin 5 and its tandem repeat

Chapter 1

Biological background

> “Biology is just applied chemistry”
>
> *- Randall Munroe*

Although one can consider both natural occuring and artificial proteins, this work mainly concerns proteins that are naturally occuring and have biological functions, or derivatives thereof. A general introduction to proteins is given, along with the class of proteins this Thesis concerns, providing the biological context these proteins operate in.

1 Proteins

Proteins are polymers, with the monomers consisting of amino acids. An amino acid in general has the structure found in Figure 1.1, with the group R differentiating between different amino acids.

In nature, there are mainly twenty different amino acids, which are covalently joined together in a linear fashion to form a protein. If there are only few amino acids joined together, then the sequence is called a peptide. The description of the sequence of amino acids in a protein is called the *primary structure*. The chain of amino acids can adopt multiple configurations, where distinct 3D-structures along a part of the chain are referred to as

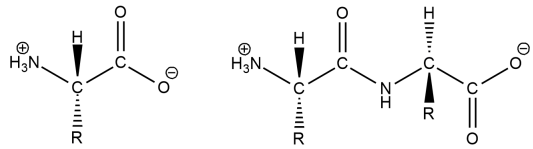


Figure 1.1: **Left:** The general structure of an amino acid. **Right:** Amino acids combining to form a (di-) peptide.

the *secondary structure*. A number of categories of secondary structure are used. The exact categorisation may vary, but the main ones are based on the hydrogen bonding pattern, which yields either helix, beta-sheet or coil structure. Another categorisation that has the advantage of being continous, rather than defining categories which some structures may not be easily categorised into, is the dihedral angles. The dihedral angles for an amino acid in a chain are three, visualized in Figure 1.2:

- ω angle, defined by the C_α atom of the previous amino acid, the carbonyl carbon of the previous amino acid, the nitrogen atom of the present amino acid, and the C_α atom of the present amino acid
- ϕ angle, defined by the carbonyl carbon of the previous amino acid, the nitrogen atom of the present amino acid, the C_α atom of the present amino acid and carbonyl carbon of the present amino acid
- ψ angle, defined by the nitrogen atom of the present amino acid, the C_α atom of the present amino acid, the carbonyl carbon of the present amino acid and the nitrogen atom of the next amino acid in the chain.

The secondary structures formed may contribute to the formation of an overall 3D-structure of the whole protein, called the *tertiary structure*. Even further, different proteins with different structures may come together to form more complex structures, called the *quartenary structure*.

An important feature of proteins is that the sequence of amino acids influences the formation of the secondary and tertiary structure. This feature has motivated the field of bioinformatics to attempt the prediction of structurual features using statistical analysis of the amino acid composition of proteins.

2 Intrinsically Disordered Proteins

A particular grouping of proteins that emerged in the early 2000s was the intrinsically disordered proteins (IDPs), which also has, historically, been denoted ”natively unfolded” or

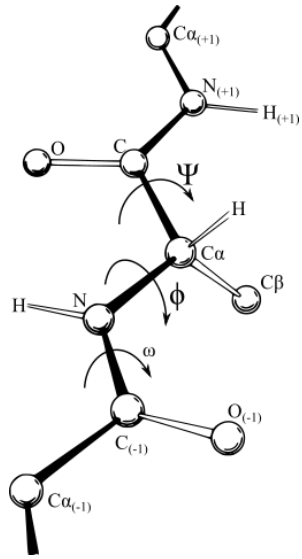


Figure 1.2: The dihedral angles ω, ϕ, and ψ in a protein chain. Taken from Wikipedia (`https://en.wikipedia.org/wiki/File:Protein_backbone_PhiPsiOmega_drawing.svg`), usage allowed under the Creative Commons Attribution 3.0 Unported licence

intrinsically unstructured. [1] The main feature of IDPs is that they do not form a distinct, singular tertiary structure. Instead, they exist as ensembles of heterogenous structures. Historically, a dogma has been that for a protein to function, a distinct structure is necessary. [2] This dogma does not apply to IDPs, a wide range of biological functions have been elucidated, despite their lack of distinct structure. Examples of these functions include transcription activation, misfolding recognition, and stress response. They have also been implicated to have roles in diseases such as Parkinson's and cancer. [3–6] However, a functionality that IDPs so far seem to be lacking is catalytic activity, they are not enzymes.[7]

IDPs are not a fringe-class of proteins. Estimates show that 25-30 % of all eukaryotic proteins are IDPs, while more than 50 % have regions in their sequence that are disordered. [8] It has also been shown that eukaryotic species on average has more disorder in their proteomes than prokaryotic organisms, though there is overlap between the groups. Likewise, organisms that change habitats have more disordered proteins than organisms that do not, like endosymbionts. [9] Bioinformatical analysis of IDPs show some amino acids to be more common in IDPs than in other proteins. Generally, charged or hydrophilic amino acids tend to be overrepresented in IDPs compared to the content globular proteins. More rigourous statistical analysis has shown that the top 5 amino acids for promoting disorder are proline, glutamic acid, serine, lysine and glutamine, all of which are either polar or charged (except for proline, which may be regarded as a special case) while the top 5 order promoting amino acids are tryptophan, phenylalanine, tyrosine, isoleucine and methionine, all regarded to be hydrophobic. [10]

Recognizing that charge is important in IDP sequences, one way to categorize proteins

suggested by Das and Pappu [11] is to consider the fraction of positive residues (f_+) and the fraction of negative residues (f_-). These measures define the fraction of charged residues (FCR = $f_+ + f_-$), and the net charge per residue (NCPR = $|f_+ - f_-|$), for which thresholds are defined to delineate different proteins into groups. This is plotted in Figure 1.3, with the position of the protein Histatin 5 shown.

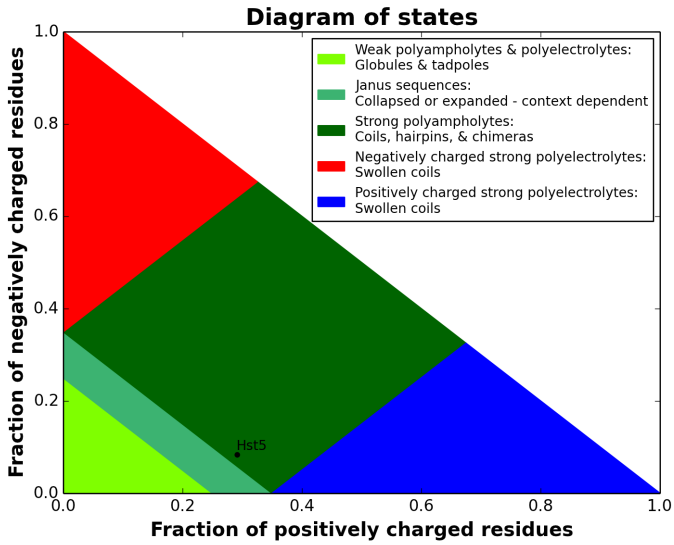


Figure 1.3: The categorization of proteins according to Das and Pappu [11], where FCR and NCPR predicts what kind of structure a protein will have, as seen in the legend. The classification of the protein Histatin 5 shown in the plot. Plot produced through the tool "CIDER". [12]

Three principle categories are found by this method: globules and tadpoles (low FCR and low NCPR), hairpins, coils and chimeras (high FCR and low NCPR) and swollen coils (high FCR and high NCPR). A fourth class, sharing features of both the first and second categories (light-green and dark green in Figure 1.3) is also considered, where the exact nature of the protein may be context dependent.

One important consequence of disorder in proteins is that they, because of their lack of distinct structure, cannot be probed using techniques like X-ray crystallography or cryogenic electron microscopy. Other techniques that can be used to study IDPs only provide average properties of IDPs. In recent years, great success has been achieved in prediction of protein structures, using the AlphaFold model. [13] However, IDPs still elude this impressive model. [14–16]

For the simpler problem of determining whether a sequence or region is disordered or not, there are several predictors available. [17–19] Using the metric of the area under the receiving operating characteristic curve, these may achieve a performance above 90 %, while accuracy, defined as the mean of specificity and sensitivity, is lower, around 80 %. [20]

3 Saliva

Saliva is 99 % water, the rest being salts, proteins and nitrogenous compounds. [21] Despite the deceptively small amount of non-water compounds, saliva has five major functions: (1) lubrication and protection, (2) acting as a buffer, (3) help maintain tooth integrity, (4) help with taste and digestion, and (5) antibacterial activity. For the lubrication function, which combats irritants, help with speech and swallowing, a group of proteins called mucins are the prime components. For the buffering activity (saliva has a pH between 6 and 7 at normal conditions), bicarbonate is the main actor, while phosphate, urea and amphoteric proteins and enzymes contribute. The buffering action also helps with tooth integrity, as development of caries has a critical pH range of 5 - 5.5. But tooth integrity is also maintained through the stabilization of calcium and phosphate salts, which may remineralize the enamel, by for example the protein statherin. Digestion is helped by the presence of amylase, which breaks down starch, and other enzymes initiate the breakdown of fat. The function most related to this Thesis is the antibactierial activity. Saliva contains molecules part of the immunesystem, such as antibodies IgA, IgG and IgM, but also the families of smaller proteins of cysteins, histatins, and proline-rich proteins. These may, for example, help aggregate bacteria, making it harder for bacteria to colonize, or inhibit specific enzymes (such as cystein-proteinase) that are important for development of disease. The histatin family is in particular also known for having antifungal properties. [22]

4 Histatin 5

The main protein of study in this Thesis is Histatin 5 (Hst5 for short), along with a derivative of Hst5. It is a member of the histatins, has a net charge of +5 and is 24 amino acids long, with a sequence as shown in Figure 1.4.

DSHAKRHHGYKRKFHEKHHSHRGY

Figure 1.4: One-letter sequence code of Hst5. Positively charged residues in red, negatively charged residues in blue and neutral residues in black.

Hst5 has been found by small angle X-ray scattering to be an IDP [23], but before the concept of IDPs was established, it was known from nuclear magnetic resonance and circular dichroism studies to lack distinct structure. [24, 25] However, it is known to form an α-helix in dimethylsulfoxide [25] and in aqueous trifluoroethanol solutions. [26] In the saliva, it is fungicidal, being particularly potent towards *Candida albicans* [27, 28] and can also bind tannins. [29, 30] Others have also investigated derivatives of Hst5, finding some of these to self-assemble and to be more potent in terms of antimicrobial properties. [31]

5 Crowding

The macromolecular concentration in cells is high, around 300 - 400 mg/mL. [32] Thus, the *in vivo* cellular environment is a crowded one, which may have impact on protein structure. However, most research investigates proteins under dilute conditions, hence a discrepancy between research data and the biological situation may arise.

One simple effect of crowding is that the available volume for a singular protein molecule becomes smaller. The effect can be rationalized by considering Le Chatelier's principle, that a system should counteract any change to the system, in this case countering the diminishing volume by minimizing the volume occupied by the protein studied. This behaviour has been found for some IDPs, with decreasing size upon crowding. [33, 34] Though, even when considering systems where this excluded volume-effect is the main consequence of crowding, the effect on protein structure is non-monotonic. [35] The situation may become even more complicated by protein-protein interactions, which are considered to be negligible in dilute systems.

Some general classes of IDPs, based on the effect of crowding, has been suggested: Foldable, non-foldable and un-foldable. [36] The foldable IDPs assume an ordered (or partially ordered) structure when present in a crowded millieu. [34, 37] Historically, when the concept of intrinsic disorder in proteins was still debated, a hypothesis was that all IDPs belonged to this class, which would render the concept irrelevant, as the disordered structure would not be present in biological settings. Non-foldable IDPs on the other hand do not exhibit any significant effect of crowding. [38–41] The last class, un-foldable IDPs, behave counter-intuitive to the excluded-effect and expand upon crowding.

These classes are not without exception, and should not be taken as absolute. An example where these classes may not be relevant is when crowding causes different sub-populations of the same protein in a solution to behave differently, where one subpopulation exhibits compaction while another subpopulation exhibits expansion. [42]

Crowding can also cause a process called liquid-liquid phase separation (LLPS). By analogy, it is similar to oil in water: one can dissolve small amounts of oil in water, but if larger amounts of oil is added, it will not dissolve but instead create another phase in the solution. This has been shown to also apply for IDPs, which at high protein concentration may create droplets in the solution. *In vivo*, such droplets have been designated membraneless organelles (MLO), providing compartmentalisation for biological functions, without the use of membranes. [43] However, a protein being an IDP is not a guarantee for LLPS, but there has been found some requirements on the IDP for LLPS, such as the IDP being trivalent. [44] Some researchers have found LLPS to be electrostatically driven, [45, 46] while others have found aromatic amino acid residues to be of particular importance for LLPS [47] and yet others find that both of these factors are important. [48] There have

been several algorithms developed for prediction of LLPS, [49] though prion-like domain predictive models have also been considered for LLPS. [50] These are considered to be "first-generation" predictors - while having predictive capability, there is still room for improvement.

Crowding may also have an impact on diffusive properties. In this context, a study comparing the diffusion rates of an IDP (α-synuclein) and a globular protein (chymotrypsin inhibitor 2) found that in dilute conditions, the smaller globular protein had a faster diffusion, while for several different crowding conditions (using both synthetic polymers and other globular proteins as crowders), the IDP had faster diffusion rate. [51] An exception to this behaviour was found when using glycerol as a crowder, where the globular protein had faster diffusion, showing that the identity of the crowder matters. Another study used a more realistic crowding environment by performing experiments inside cells. [52] Also this study found an IDP to diffuse faster in crowded environments than a globular protein of similar dimensions.

Crowding is not just a concern in terms of how proteins behave *in vivo*. A medical application is the formulation of protein therapeutics, where a high protein concentration is used to achieve correct dosage when administration is done via the subcutaneous route, which is more appealing as it, for example, can be used at home. [53, 54] High protein concentration may also be pursued to decrease the frequency at which the therapeutic needs to be injected. [55] The challenges that appear in this context is limited solubility of the protein, the possibility of protein aggregation and increased viscosity which may have negative impact on large-scale manufacturing processes, the deliverability of the protein and the shelf-life of the therapeutic.

Chapter 2

Theory

Theories provide frameworks for interpretation of the world, rationalizing methods and sometimes predictions of future results. Here, the underlying rationalization for molecular simulation, a description of forces at the molecular level, and a description of polymer behaviour is presented.

1 Statistical thermodynamics

Thermodynamic properties can be observed or measured with the naked eye - a so called macroscopic scale. Observations on this scale are the averages of many particles, on the order of, for example, 10^{20}. Statistical thermodynamics is used to connect these averages with properties of molecules.

A full description of the foundations of statistical thermodynamics can be found in, for example, the book "An Introduction to Statistical Thermodynamics" by Hill [56]. Here, a brief overview of the theory is presented.

The concept of *ensembles* is a pillar in statistical thermodynamics. An ensemble is a mental collection of many systems. Each system considered to have the same general thermodynamical properties, but can vary on the microscopical level, as many different states may yield the same thermodynamic state. Depending on what the set thermodynamical variables are, the type of the ensemble is determined. Most relevant in this Thesis is the ensemble where the number of particles N, the volume V, and the temperature T are set constant. This ensemble is referred to as the *canonical ensemble*, or short-hand written as the NVT-ensemble. Having both N and V constant would be considered a closed sys-

tem. Other common ensembles are the *microcanonical ensemble*, where the energy E is set constant instead of the temperature, and the *grand canonical ensemble*, where the chemical potential μ is constant, along with volume and temperature.

A postulate concerning these ensembles states that

> *the (long) time average of a mechanical variable M in the thermodynamic system of interest is equal to the ensemble average of M, in the limit of the size of the ensemble approaching infinity, provided that the systems of the ensemble replicate the thermodynamic state and environment of the actual system of interest.*

Thus, instead of measuring the time-averaged result of a singular system, we can instead take an average over a large number systems in any given moment.

The systems in the ensemble can be found to be in many different states, or configurations. Enumerating each state, one can count in how many systems a state is found, forming a distribution. Combinatorically, for a given distribution of n_1 states being in state 1, n_2 states being in state 2, etc., the number of states, denoted Ω, is

$$\Omega = \frac{(n_1 + n_2 + n_3 + ...)!}{n_1! n_2 n_3! ...} \tag{2.1}$$

This applies for a given distribution.

To compute the average of a (mechanical) variable, a summation over all the states need to be done, summing the probability for a specific state and the value of the variable in that state. Taking pressure (denoted p) as example, we get

$$< p >= \sum_j p_j P_j \tag{2.2}$$

The probability of a state, the number of systems in the state of interest divided by the total number of systems, can be determined if the distribution is known. It can be reasoned that the most probable distribution will dominate, and the most probable distribution is the one that has the largest number of states Ω. The task becomes one of maximisation of Ω.

The logarithm of Eq. 2.1, which is more convenient, can be approximated with Stirling's approximation. Then, the maximum of Ω is found by using the method of Lagrange undetermined multipliers, with the trivial requirements that the energy of each state multiplied by the number of such states summed yields the total energy and that summing all

systems in states 1, 2, 3 ... yields the total number of systems. Then, the probability is found to be

$$P_j = \frac{e^{-\beta E_j}}{\sum_i e^{-\beta E_j}} \tag{2.3}$$

where $\beta = 1/(k_B \cdot T)$ with k_B being the Boltzmann constant, E_j the energy of state j and the factor $e^{-\beta E_j}$ is called the Boltzmann factor. A consequence of this expression is that it is lower probability to find states with high energy than systems with low energy.

It can further be shown that the denominator in Eq. 2.3, which is the definition of the partition function Q, can be related to all thermodynamic properties of interest, via the Helmholtz free energy A:

$$A = -k_B T ln(Q) \tag{2.4}$$

The treatment so far is general, applying to both quantum-systems and classical systems. However, for classical systems, an approximation is usually made where the discrete states are considered to be similar enough so they can be considered continous. Then, the sum in the expression of Q can be changed to an integral, though with a correcting pre-factor:

$$Q_{classical} = \frac{1}{N! h^{3N}} \int_S e^{-\beta E(s)} \tag{2.5}$$

where h is the Planck constant, S is the phase space and s a point in phase space. Unfortunately, this cannot be computed analytically, but numerical methods are necessary.

2 Intermolecular forces

2.1 Electrostatics

Electrostatics concerns the forces that charged particles have upon each other. The central equation, Coulomb's law, calculates the force between two charges:

$$F = \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r r^2} \tag{2.6}$$

where q_1,q_2 are the charges, ϵ_0 is the vacuum permittivity (a constant set to $8.854 \cdot 10^{-12} Fm^{-1}$), r is the distance between the charges and ϵ_r is the relative permittivity. The relative per-

mittivity introduces any medium between the charges implicitly, and will thus depend on what the medium is. Polar solvents will yield larger relative permittivity, screening the interaction between the charges. Coulomb's law can also be expressed in terms of energy, by integrating over two distances - the energy required to move the charges at a distance r_1 to a distance r_2. A reference state can be set by setting the initial distance r_1 to infinity, why the interaction energy between two charged particles in a medium becomes

$$E = \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r r} \tag{2.7}$$

The interaction energy can be seen as a charge, moving from an infinite distance, up against an electrostatic potential generated by the second charge. The electrostatic potential can be obtained by removing either q_1 or q_2 in the right-handside of Equation 2.7.

Additional screening can be provided by salt ions in a solution. These can also be treated implicitly, through the Debye-Hückel equation:

$$V(r) = \psi_0 e^{-\kappa r} = \frac{q}{4\pi\epsilon_0\epsilon_r r} e^{-\kappa r} \tag{2.8}$$

where V is the potential, ψ is the potential generated by the charge, which is identified as the expression for electrostatic potential from Coulomb's law, r is the distance and κ is the inverse Debye screening length. The screening length determines the decrease in interaction between any charges, and is given by

$$\kappa = \sqrt{\frac{2Ie^2}{\epsilon_0\epsilon_r RT}} \tag{2.9}$$

where e is the elementary charge, T is the temperature and R is the gas constant. I is the molar ionic strength of the solution, given by

$$I = \sum_i (c_+ z_+^2 + c_- z_-^2)/2 \tag{2.10}$$

where c is the cationic(+) or anionic(-) concentration and z is the valency for all the salts i in the solution. The Debye-Hückel equation is valid for small potentials and for salts that fully dissociate in solution. Ions are also assumed to be point-charges.

2.2 Dipoles

Atoms and molecules may have a neutral net charge, but they still consist of positive nuclei and negative electrons. The electrons distribute according to the laws of quantum mechanics, but may not distribute evenly. This gives rise to partial charges in molecules, which can be described as multipoles. The simplest non-evenly distributed multipole is the dipole, where two partial charges (of different sign, but same magnitude) can be formed from the distribution of electrons. The formula describing a dipole is

$$\vec{\mu} = ql \tag{2.11}$$

where q is the magnitude of the charges and l is the distance between the charges. As indicated in Eq. 2.11, dipoles are described as vectors, i.e. they are directional.

Dipoles can be permanent or induced. A permanent dipole exists due to the difference in ability to attract electrons (electronegativity) between two neighbouring atoms in a molecule. A permanent dipole also requires the molecule to be asymmetric, otherwise other atoms may cancel out the effect.

An induced dipole on the other hand is created from the influence of an external electrical field, which is supplied by, for example, a nearby charge or permanent dipole.

2.3 van der Waals forces

van der Waals forces is an umbrella term for interactions that have distance dependence of r^{-6}. They are all dipole-interactions, but differ in the kind of dipoles that interact. These interactions are called *Keesom interaction*, *Debye interaction*, and *London-dispersion interactions*. The Keesom interaction is the interaction between two permanent dipoles. The Debye interaction is the interaction between a permanent dipole and an induced dipole, produced by the permanent dipole. London dispersion interactions are instantaneous dipoles that induce dipoles, from which there is an interaction. The instantaneous dipole forms from the instantaneous distributions of electrons in molecules, which may momentarily be assymetric. The effect is quantum-mechanical in origin.

Out of the three interactions contributing to van der Waals forces, London dispersion interactions is the only one present in all molecular systems. It should also be noted that this interaction is not additive; two molecules interacting through London dispersion interactions will be affected by other molecules in their vicinity.

2.4 Hydrogen Bonds

A hydrogen bond is a special case of a dipole-dipole interaction. A dipole constituted from a hydrogen atom and either an oxygen, nitrogen, fluorine or a chlorine atom will be paticularly strong, since the difference of electronnegativity is large. The hydrogen bonded to the other atom (denoted X), can interact with another atom being oxygen, nitrogen, fluorine, or chlorine (denoted Y) which creates an interaction about a magnitude stronger than a van der Waals interaction, though not as strong as a covalent bond. The X-H part of interaction is referred to as the hydrogen bond donor, while the Y part of the interaction is referred to as the hydrogen bond acceptor.

2.5 Steric repulsion

Due to quantum mechanical effects, electron clouds of separate atoms may not overlap. This has the effect that there is a limit of how close two atoms may approach each other. This is called steric repulsion. Approximatively, once can consider this limit "strictly", like putting two billiard balls next to each other - unless a forced strong enough to break the balls is used, they will not come closer to each other. This approximation refers atoms as *hard spheres*. Alternatively, the repulsion can be modelled with a powerlaw. A common example of a powerlaw is r^{-12}, as is done in the Lennard-Jones interaction.

2.6 Entropic effect

Entropy refers to the number of configurations a system can achieve, and the relative probability of these. If two molecules approach each other, some of these configurations will be restricted. This leads to a reduction in entropy, which, if no other forces compensate for the loss of entropy (which relates to energy via the temperature), will lead to a repulsive effect. This should not be confused with the steric repulsion.

3 Polymer theory

The interactions of a singular polymer chain may either be between atoms within the polymer chain (intramolecular interactions) or with solvent atoms, considering a very dilute solution so that no polymer chain interacts with any other polymer chain. The relative strength of (all) intramolecular interactions and (all) solvent interactions can be summed up in a parameter w, which is the difference in interaction energy between all monomer-monomer particles and all the monomer-solvent particles. To account for the effect of entropy, w is used to form the parameter χ, defined as

$$\chi = w/RT \tag{2.12}$$

where R is the gas constant and T is the temperature. For different values of χ, three regions are identified:

- $\chi < 0.5$: Compared with thermal fluctuations or solvent interactions, intramolecular interactions are weak. This is referred to as the polymer being in a *good solvent*.
- $\chi > 0.5$: Intramolecular interactions are relatively strong. This is referred to as the polymer being in a *bad solvent*.
- $\chi = 0.5$: Neither interaction dominates. No interaction would seemingly affect the polymer chain. This condition is called a theta (θ) condition.

The parameter of χ and the resulting regime of χ will affect the structure of the polymer. In order to describe this influence, a measure of structure is needed. One such measure is the *radius of gyration*, R_g, which, assuming all monomers are the same (homopolymer), is defined as

$$R_g = \frac{\sum_{i=1}^{p} |\vec{r_i} - \vec{r_{CM}}|^2}{N_p} \tag{2.13}$$

where $\vec{r_i}$ is the position of monomer i and $\vec{r_{CM}}$ is the center-of-mass position. N_p is the number of monomers in the chain. Another useful metric of a structure is the end-to-end distance, which is the distance between the first amino acid in the sequence (the N-terminal) and the last amino acid in the sequence (the C-terminal):

$$R_{ee} = \sqrt{< |\vec{r_1} - \vec{r_p}|^2 >} \tag{2.14}$$

where the angular brackets denote an average over all conformers considered. Another measure for structure is the hydrodynamic radius (R_h), which can be defined in terms of diffusion. Considering a spherical particle, the hydrodynamic radius is the radius of the sphere that diffuses through a medium according to the Stokes-Einstein equation,

$$R_h = \frac{k_B T}{6\pi\eta D} \tag{2.15}$$

where k_B is the Boltzmann constant, T is the temperature, η is the viscosity of the medium and D is the diffusion coefficient. For any geometry not being a sphere, the hydrodynamic radius is an "effective radius", i.e. the radius it would have if it had been a sphere.

3.1 Scaling laws

At theta conditions, where there are (effectively) no interactions, a model to describe the overall size of polymers is the *random walk*, also called *freely jointed chain*. In a random walk, from the start position, a random direction is chosen and a distance l is travelled in that direction. From the new position, a new direction is randomly chosen and the same distance l is travelled in that direction. This is repeated, up to a specified number of steps. The distance between the starting position and the end position can then be, on average, computed as a function of the square root of the number of steps taken and the length of the step taken. Adapting this model to polymer structures, the step length is identified as the monomer size and the number of steps the number of monomers. A complication is found for stiff chain segments, where the step length would be identified as a certain number of monomers, but this can later be accounted for in a pre-factor. Thus, a polymer at theta-conditions, following a random-walk model, would have the following scaling law:

$$R_g = \alpha N_p^{0.5} \tag{2.16}$$

where α is a prefactor and N_p the number of monomers. However, this model allows for the path taken to cross itself. Molecules have steric repulsions which do not allow this, therefire a modification of the random walk into the self-avoiding random walk (SARW) is introduced. The SARW is less compact than the random walk, and is described by the equation

$$R_g = \alpha N_p^{0.6} \tag{2.17}$$

which then is the expected size of a polymer at theta-conditions.

If the polymer is in a bad solvent, the polymer would minimize the surface area with the solvent, becoming a more compact globule. Such a polymer would instead follow the scaling law

$$R_g = \alpha N_p^{0.33} \tag{2.18}$$

One should consider that all Equations 2.16 - 2.18 applies for polymers with large numbers of monomers. Smaller polymers may need corrections. As well, these scaling laws are approximate for heteropolymers (polymers with different kinds of monomers, such as proteins), as the size of each monomer may be different - an effective monomer size would need to be considered.

For diffusion, a scaling law has been suggested by Augé *et al.* [57], which does not consider the number of monomers in the chain, but the mass M and the fractal dimension d_F,

$$log(D) = -\frac{1}{d_F} log(M) + log(C). \tag{2.19}$$

where C is a constant which depends on the molecular "family" that is considered and needs to be parameterized. Scaling laws have also been developed for the hydrodynamic radius, with various factors considered. [58] Only considering the number of residues, for an IDP a scaling law suggested is Eq. 2.20:

$$R_h = R_0 N^{\nu} \tag{2.20}$$

where R_0 and ν are constants. A more detailed scaling law taking into account the fraction of prolines in the sequence (f_{PRO}) and the net charge (q) is

$$R_h = (1.24 \cdot f_{PRO} + 0.904) \cdot (0.00759 \cdot |q| + 0.963) \cdot 2.49 \cdot N^{0.509} \tag{2.21}$$

as parameterized by Marsh and Forman-Kay. [58]

Chapter 3

Molecular Simulation

There are several ways to perform a molecular simulation. In this work, two approaches have mainly been employed: Monte Carlo and Molecular Dynamics.

1 Monte Carlo

The first ever Monte Carlo (MC) molecular simulation was performed 1953 [59] on the Los Alamos MANIAC computer. The method can be adapted for more than just chemical simulation; many optimization problems in many different fields may be approached. A few examples of this is how COVID 19 spread [60], risk assesments for nuclear power-plants [61], how to choose a portfolio for investing [62], and predict air travel demand. [63] For molecular simulation, the objective is to sample states (configurations) so that the probability found in Eq. 2.3 is estimated. Sampling all possible states is computationally impossible, but the idea in MC is that some states have such low probability that they can be ignored, so the algorithm aims to only sample states with high probability. As found in Eq. 2.3, the probability is proportional to the Boltzmann factor. The basic MC algorithm is described below:

1. Generate an initial configuration
2. For a number of steps (the duration of the simulation):
   (a) Select particle at random, calculate its energy in phase space, E(s)
   (b) Displace the particle randomly, $s' = s + \Delta$
   (c) Compute the new energy in phase space, E(s')
      - Accept the move if E(s') < E(s)
      - If E(s') > E(s), generate a random number $X \in [0, 1]$
      - If $X < e^{-[E(s')-E(s)]/k_B T}$, accept the move
      - Otherwise, reject the move and return the particle to previous position in phase space

1.1 Trial moves

The displacement in the MC algorithm can include many types of changes to the configuration of the system. In particular, these so called Monte Carlo moves can be unphysical, which is one of the reasons dynamical properties cannot be estimated using MC. There are some restrictions in regards how these moves should be constructed; mainly the condition of *detailed balance*, which is discussed in detail in Ref. [64]. Briefly, detailed balance is imposed to fulfil the requirement of moves not changing the distribution once the equilibrium distribution has been attained. Detailed balance is in this context a stronger than necessary requirement, being the condition of

$$P(s)\pi(s \to s') = P(s')\pi(s' \to s) \tag{3.1}$$

where $\pi(s \to s')$ is the transition probability of going from state s to state s'. As examples of moves, consider the following MC moves found in the bead-necklace model used in this work:

Single particle translation randomly takes a singular particle and moves it a distance (in any direction) away from its current position.

Pivot rotation is a rotation of a (polymer) chain about an axis, decided upon by random selection of a bond in the chain.

Chain translation randomly takes a whole (polymer) chain and moves it a distance, while keeping the chain rigid.

Slithering move, also called reptation, randomly chooses a chain and takes one of its end-particles and displaces it. The rest of the chain, while keeping the chain rigid, is moved along with the displaced end-particle.

Size of a displacement

Some of the above-mentioned moves require a parameter, that determines how dramatic the change is. This parameter should not be too small nor too large. If it is too small, only small changes happen, which can both increases simulation time and make the crossing of energy barriers harder, possibly only sampling local energy minima. A too large parameter may lead to larger differences in energy, making trial moves less likely to be accepted, increasing simulation time. This is a balance that depends on both the system of study and the interaction potential chosen. There exists methods where the displacement size is changed during the simulation to make the simulation more efficient, [65–67], but these are not considered in this Thesis.

2 Molecular Dynamics

Molecular Dynamics (MD) is, compared with MC, a younger algorithm for chemical simulation - the first MD simulation was done in 1957. [68] A distinct advantage of the method over MC is that dynamical properties can be obtained. The method uses Newton's equations of motion. After an initial setup of the system in a simulation box (usually including energy minimization to ascertain initial forces are not too large), velocities are randomly assigned to the particles, but are shifted so that the velocity center of mass is zero. The velocities are scaled to fit with the initially set temperature, as velocity of particles and temperature are connected by the equipartition theorem:

$$T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{k_b N_f} \tag{3.2}$$

where $T(t)$ is the temperature as a function of time, m_i is the mass of particle i, v_i is the velocity of particle i, k_B is the Boltzmann constant and N_f is the number of degrees of freedom (for a fized total momentum, this is 3N-3, N being the number of particles). After this initial setup, the simulation can start. Three general operations are done throughout the simulation: the calculation of all forces, the integration of Newton's equations and registration of variables for computation of any sought-after average properties of the system.

2.1 Verlet algorithm

In this work, the Leap-frog algorithm [69] has been used for the integration of the Newtonian equations. This is a variation of the Verlet algorithm, which is derived by considering the Taylor expansion of the coordinate of a particle, in regards to time. For particle moving in one dimension, the next position after a time Δt, while currently occupying coordinate r(t), is

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4) \tag{3.3}$$

where v is velocity, f is the force, m the mass and O the order of the error. The equation also holds for going back in time; i.e. the previous coordinate can be computed by

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4) \tag{3.4}$$

Adding these equations together and rearranging, omitting the error, the next coordinates can be computed by

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 \tag{3.5}$$

which is one of the simplest algorithms, but still achieves an error of $O(\Delta t^4)$. In order to achieve higher precision, one can take into account higher-order terms in the Taylor expansion. These may however have other drawbacks (see Ref. [69]) and are not considered in this Thesis.

2.2 Bond constraints

The time step, Δt, need to be set as to not introduce a too high error as shown by the above discussion, while also considering that smaller time steps translates to higher computational load, for a given simulation time. Additionaly, the time step should consider the highest frequency motion in the system, to prevent the simulation becoming unstable. This highest frequency motion in the system is usually bond vibrations, which commonly has put the time step to be about 1 fs. [70]

This limitation can be overcome by using a constraint for the bond vibrations. An example of such a constraint is the LINCS algorithm. [71] In this algorithm, the new positions of two particles bonded together are first computed as above, but this position is then

corrected according to a projection to the old direction of the bonds, whereafter the bond length (and thereby the positions) is adjusted in the new bond direction to retain target bond length.

2.3 Keeping temperature constant

As different ensembles could be aimed at in a simulation, variables may need to be kept constant. To keep the temperature constant, which is needed in the canonical (NVT) ensemble considered in this Thesis, one could scales the velocities in every step to be in accordance with the equipartition teorem (Eq. 3.2), but with this restrained, there would not be any fluctuation of temperature and the kinetic energy would be constant, which would not achieve a true canonical ensemble. There are several algorithms that can be used, but in this work, the velocity-rescale thermostat [72] is used. The velocities are still scaled by a constant, but it is done at a specified interval and the scaling constant is arrived at differently. The scaling constant is found by taking a kinetic energy that is stochastically chosen from the kinetic energy distribution, as imposed by the canonical ensemble.

3 Interaction potentials

For both MC and MD simulations, an interaction potential, describing the energies of a system, is needed. In the case of MD, the interaction potential is usually referred to as a force field, as the force (the derivative of the interaction energy in regards to distance) on particles is used in the simulation. For exact descriptions of these energies, a quantum mechanical treatment is necessary, which requires a computational cost only allowing the study of small systems and limited timescales. Instead, approximations are used to enable the study of larger system over large timescales.

An important commonly used approximation is pairwise additivity, which means that the interaction between any two particles can be considered independent of any other particles and the total interaction on any singular particle is found by summing all interactions with other particles pairwise. This also reduces the computational complexity to $O(N^2)$. A related approximation is that electrons are considered to be in their ground states, since the treatment of electron distributions would require a quantum mechanical treatment. This is a fairly reasonable approximation, as the difference in mass between an electron and nuclei (say a single proton) is large ($9.1{\cdot}10^{-31}$ vs. $1.7{\cdot}10^{-27}$), so an electron can respond comparatively instantly to a change in nuclei displacement. [73]

An general example of a commonly used force field for proteins (the AMBER force field, [74] a variation of which is applied in this Thesis) is found in Eq. 3.6:

$$E_{total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedral} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] + \sum_{i<j}[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}}] \tag{3.6}$$

where r_{eq}, θ_{eq} are bonded parameters, V_n, γ are parameters for the dihedral, and q_i, q_j are partial charges forming the Coulombic forces, which also require parametrization. The first two terms in the last sum form the Lennard Jones (LJ) potential, where the first part $(\frac{A_{ij}}{r_{ij}^{12}})$ approximates steric repulsion, while the second part $(\frac{B_{ij}}{r_{ij}^{6}})$ approximates all van der Waals forces. The LJ potential is useful as the distance dependence of the steric repulsion is the square of the distance dependence of the van der Waals interaction, which is computationally faster than having non-related distance-dependencies between the two, as they would need to be computed separately. The LJ potential may also be expressed as

$$E_{LJ}(r) = 4\epsilon\left((\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}\right) \tag{3.7}$$

The parameters can be deduced by fitting to quantum mechanical calculations or by fitting to experimental data. When two particles i and j with different LJ paramters interact with each other, a rule for how the parameters should be combined need to be used. Two examples of combination rules are the geometric mean, $A_{ij} = (A_{ii}A_{jj})^{1/2}$ and $B_{ij} = (B_{ii}B_{jj})^{1/2}$, and the Lorentz-Bertelot rule [75, 76],

$$\begin{aligned} \sigma_{ij} &= \frac{\sigma_{ii} + \sigma_{jj}}{2} \\ \epsilon_{ij} &= (\epsilon_{ii}\epsilon_{jj})^{1/2} \end{aligned} \tag{3.8}$$

3.1 Force field development

The improvement of force fields is an active research field, with a fairly long development history. Some of the most widely used families of force fields (such as AMBER, CHARMM, GROMOS, and OPLS-AA) has been developed since the 1980s or 1990s. [74, 77–79] After the concept of IDPs matured throughout the 2000s, simulations of IDPs were benchmarked against experimental data. [80–82] These showed in particular that many force fields yielded too compact structures for IDPs, prompting further development.

Two methods of improvements have been widely applied: Optimization of dihedral angle parameters and changing water-protein interactions. [83, 84] Some examples of force fields improved by changing dihedral angle parameters are ff03*, ff99SB*, OPLS-AA/M, OPLS3, and CHARMM22*. [85–88] A variation on the dihedral angle optimization scheme should be noted, the grid-based energy correction map (CMAP). [89] In CMAP, the dihedral angles are divided into bins and for each bin a conformational free energy is computed,

$$\Delta G_i = -RTln\left(\frac{N_i}{N_{max}}\right) \tag{3.9}$$

where i is the bin index, N_i is the number of dihedral data in bin i and N_{max} is the total number of dihedral data. These are computed from both the force field to be corrected, and a reference database. The difference in conformational free energy between these is then incorporated into the force field, interpolating the energy for values of the dihedral angles not having the angles of the center of the bin. This correction was first used in the development of the CHARMM27 force field. Additional iterations of the same force field, during which was CMAP with NMR data as basis and using CMAP to some part in a residue specific way (and additional changes) eventually yielded CHARMM36m, a force field claimed to be "balanced" considering the simulation of both folded proteins and IDPs. [90] Another force field using this scheme in a residue-specific manner, while also only considering IDPs, is CHARMM36IDPSFF. [91]

The second widely used correction, the changing of water-protein interaction parameters, most often changes the LJ-parameters of the force field. Examples of such force fields are ff03ws, [82] which circumvents the Lorentz-Bertelot rule to have specifically interactions between water and protein changed, while keeping parameters themselves unchanged, and a99SB-*disp*, [92] which changed the water model parameters to change water-protein interactions (among other changes).

Care need to be taken when choosing a force field. For example, a newer force field is not necessarily better performing than an older, depending on the observables of interest, as shown by comparing CHARMM22* with CHARMM36m. [83] Not only need one to consider what observables that are of interest, but also the system - while some force fields try to encompass all proteins (such as a99SB-*disp* and CHARMM36m) but can have varying performance for different systems, others are more specific towards a certain class of proteins (such as CHARMM36IDPSFF).

3.2 Water models

Unlike proteins, which is a large class of molecules, which in humans alone encompass anywhere between 20 000 and 6 million members [93] (including different post-translational

modifications), water is a single compound, with a structure that has been investigated since the 1930s. [94, 95] Despite this comparable simplicity, there was in 2002 more than 40 different models of water. [96] Among other reasons is to have models that are as simple as possible, while retaining the properties of interest. Perhaps among the most simple models possible is the use of an implicit water model, where individual water molecules are not considered, but all water molecules are treated as a continuum. This yields computational efficiency, but it is a great simplification. Benefits of implicit solvents are listed by for example Onufriev and Case. [97]

Considering explicit water models, one of the most widely used family of water models is the Transferable Intermolecular Potential N-Point model (TIPNP, where N denotes the number of points the model uses), which are rigid, fixed-charge models. [98] Two important members of this family is TIP3P and TIP4P. TIP3P have three points in a v-shaped pattern at an angle of 104.52 °, where the end-points represent the H-atoms and the middle point represent the oxygen atom. Each hydrogen is assigned a charge of +0.417, and the oxygen a charge of -0.834. Interactions with other molecules is through Coulombic interactions with the charges, and a LJ-interaction for the oxygen atom only. TIP4P on the other hand introduces a fourth point, located on a bisecting line of the HOH angle, towards the hydrogens. At this point, the oxygen charge is moved, while the LJ-calculation is still based on the original oxygen position. The properties obtained from these water models are highly sensitive to parameterization, [99] and many variations exist. As well, the combination of a particular water model with a particular protein force field impacts performance. For instance, using the AMBER ff99SB-ILDN force field with a TIP3P water model has been found to yield too compact structures of Hst5, [80] while instead using a variation of TIP4P called TIP4P-D [100] yielded results more in line with experimental data. [101] Atomistic simulations with explicit solvent presented in this Thesis have been using the TIP4P-D water model and the a99SB-*disp* water model, as they, in combination with AMBER ff99SB-ILDN and a99SB-*disp* protein force field, respectively, have shown to have fair performance for Hst5.

3.3 The interaction potential in the bead-necklace model

In this work, a bead-necklace model is frequently applied using the MC simulation method. The interaction potential used is the following:

$$U_{total} = U_{bonded} + U_{non-bonded} = U_{bonded} + U_{hs} + U_{DH} + U_{short}, \tag{3.10}$$

with the bonded harmonic spring given by:

$$U_{bonded} = \sum_{i=1}^{N-1} \frac{k_{bond}}{2}(r_{i,i+1} - r_0)^2, \tag{3.11}$$

where the sum loops over all the monomers. Here, $r_{i,i+1}$ denotes the distance between two connected monomers, r_0 is the equilibrium distance set to 4.1 Å, and k_{bond} is the spring force constant set to 0.4 N/m. A hard-sphere potential is used to mimic the steric repulsion, defined as

$$u_{ij}^{hs}(r_{ij}) = \begin{cases} 0, r_{ij} > R_i + R_j \\ \infty, r_{ij} < R_i + R_j \end{cases}, \tag{3.12}$$

where R_i, and R_j are the radii of the beads. All beads are here given the same radius of 2 Å. All pairs of particles i, j (i $\neq$ j) are considered, with a total hardsphere-interaction summed to yield U_{hs}. An extended Debye-Hückel potential is used, according to

$$U_{DH} = \sum_{ij} u_{ij}^{DH} = \sum_{i<j} \frac{Z_i Z_j e^2 exp[-\kappa(r_{ij} - (R_i + R_j))]}{4\epsilon_0 \epsilon_r (1 + \kappa R_i)(1 + \kappa R_j) r_{ij}}, \tag{3.13}$$

where e is the elementary charge, Z is any integer, positive or negative, corresponding to the number of charges of a given amino acid, κ is the inverse Debye screening length, ϵ_0 is the vacuum permittivity, and ϵ_r is the dielectric constant for water. In this case, Z is either 1 or -1, depending on the charge of the amino acid; neutral particles (Z=0) are not treated by the Debye-Hückel potential. Salinity is treated via the inverse Debye screening length, as defined in the Theory part. To determine the charge of each bead (amino acid), a method by Kurut *et al.* [102] was used, where a Monte Carlo simulation is performed with a titration step, which adds a protonation term to the interaction potential,

$$\beta U_{protonation} = \sum_{i=1}^{N_p} (pK_{a,i} - pH) ln 10 \tag{3.14}$$

where N_p is the number of titratable amino acids in the sequence, and $pK_{a,i}$ is the intrinsic acid dissociation constant for amino acid i, as found by Nozaki and Tanford. [103] van der Waals forces are treated with a short-ranged potential:

$$U_{short} = -\sum_{i<j} \frac{\epsilon}{r^6}, \tag{3.15}$$

where ϵ sets the strength of the interactions due to polarisability. Here ϵ was set to $0.6 \cdot 10^4 kJ$ Å^6/mol to achieve an attractive potential of 0.6 kT at closest contact. For short range interactions, all beads are treated identically.

4 Simulation technicalities common to both MC and MD

4.1 Simulation box

In the simulation, a box defining the boundaries of the simulation volume is used. To assure that particles are simulated as in bulk, what happens when particles approach the boundaries need to be considered, otherwise the system at the edge may exhibit effects not present in the middle of the box. This is particularly important if the box is small with a small number of particles, which then would have a large fraction of particles at the edges. To avoid such behaviour, a common algorithm used is periodic boundary conditions (PBC). This imagines the simulation box to be one cell surrounded by other, perfectly identical cells. Particles in the cell would interact with particles in neighbouring cells - including themselves, as every cell is an identical copy of the cell of interest. To avoid particles interaction with themselves, a cut-off for the interactions is used. Practically, when a particle diffuse/translates across a box boundary, it is put on the other side of the box when using PBC.

Despite the use of tricks to assure bulk properties are achieved in the simulation, there are properties that are dependent on the box size. For example, the longest wavelength of fluctuations will be the size of a box side length. Therefore, one has to consider if the box is adequately large to be able to consider the properties of interest, while keeping in mind that a larger box requires a greater number of particles to simulate, if a particle concentration should remain at a specific value.

4.2 Interaction truncation

van der Waals interactions and steric repulsions are usually treated with potentials where the interaction decreases fairly quickly with distance, so that the contribution to the energy is small at large distances. A cut-off can then be used, where only the interaction of particles within a specified distance (the cut-off) are considered. Interactions with far-away particles are either ignored or dealt with through a correction, which would not treat the remaining particles explicitly. The procedure of using cut-offs also decreases the number of computations needed, decreasing computer power needed.

However, in the case of electrostatic interactions, the interaction does not decrease as

quickly with distance, so long-range interactions contribute significantly to the overall interaction energy, so cut-off need to be as large as possible (half the box-side length, to avoid previously stated problems with PBC). As an alternative, the algorithm of *Particle Mesh Ewald* (PME) [104] can be used (there are other algorithms achieving the same effect, but PME is one of the most popular). PME is a faster variation of the Ewald method [105], which divides the calculation into two parts: Short range interactions are computed in real space, while long-range interactions are computed via a Fourier transform.

Neighbor list

If a potential with a cut-off is used, it is possible to save computer power if the system is large (the simulation box need to be larger than the cut-off distance). This can be achieved with either the Verlet-list method [106] or the cell-list method. [107] The Verlet list is a list of the particles which will actually be interacting with a particle i, due to the cut-off distances being exceeded for other particles. Even if truncation is used, the distance bewteen any particle and all others need to be computed, amounting to N(N-1)/2 pair distance computations. In the Verlet list scheme, a list of particles within a distance r_v, larger than the cut-off distance r_c, is created. If, in the next step, no particle has a displacement greater than r_v - r_c, then only the particles in the Verlet list need to be considered, skipping all particles not in the list. If a particle has a displacement larger than r_v - r_c, then the Verlet list need to be updated, but this is not needed for every interaction calculation, thus decreasing the number of computations needed.

In the cell-list method, the whole simulation box is divided into cells, each with a side-size equal to or slightly larger than the cut-off distance r_c. For each particle, the pair distance computation now only need to be considered for particles in the cell it is currently in, and the neighboring cells. In particular, the updating of the cell-list scales with N.

Which method is more efficient to use depends on the system, in particular the system size. A discussion on the efficiency of either method (and the combination of both methods) can be found in the book by Frenkel and Smit. [64]

4.3 Convergence

In both MC and MD, the aim is to sample states (configurations) so that correct probability for different states are achieved, from which one can compute average properties of a system. An important question that arises in this context is how long does a simulation need to run in order to obtain correct probabilities. A simulation is said to have reached *convergnce* if correct probabilities have been estimated. The problem of convergence is also known as the sampling problem. For an infinitely long simulation, the value of any property should reach

a constant value. Since computer resources are not infinite, one instead have to rephrase this requirement so that the simulation has converged if a variable is seemingly at a constant value. Strictly, convergence is impossible to prove, but can be argued for any property by plotting the property against simulation time, where one can visually inspect the behaviour and make a judgement regarding whether a constant value has, within a specified error, been achieved.

A problem in regard to convergence is that a simulation may only sample a subpopulation of configurations, as there can be an energy barrier between subpopulations of configurations that need to be crossed for proper sampling. This is referred to as the simulation being trapped in a local energy minima. There are special techniques to overcome this problem, such as metadynamics or simulated annealing, which are not considered in this Thesis. There are several statistical methods to help judge the convergence of a simulation (or rather, a single observable in a simulation), but each also have weaknesses. [108] Two common procedures should be mentioned: correlation-time analysis and block averaging.

Correlation-time analysis

The (auto) correlation-time is described as the time that it takes for a process to "forget" which values it had earlier, more strictly how fast correlation between datapoints different in time is lost. The correlation-time is computed from the autocorrelation function $c(t')$, which for a function $f(t)$ is

$$c(t') = \frac{(1/N)\sum^{N-(t'/\Delta t)}[f(t) - \langle f \rangle][f(t+t') - \langle f \rangle]}{\sigma^2} \tag{3.16}$$

where N is the number of datapoints, Δt is the time step of the simulation, $\langle f \rangle$ is the mean of the function f and σ^2 is the averaged squared deviation from the mean. Note that a more compact version of this formula is found in the section about dynamic light scattering. The correlation-time is computed via

$$\tau = \int_0^\infty dt' c(t') \tag{3.17}$$

The correlation-time can be used to find the number of statistically independent data points in the simulation. It has been argued that one should have at least 20 such independent data points to have a reliable estimate of the property investigated. [108] Weaknesses of this procedure are not accounting for the full statistical information in the trajectory (only using $N-(t'/\Delta t)$ frames) and being less reliable when processes on a slower timescale comes into play in the simulation.

Block averaging

In block averaging, the trajectory is split into M pieces, each n long. The average value of the function of interest for each M is computed, out of which a standard deviation is computed, σ_n, and the block standard error (BSE) is computed

$$BSE(n) = \frac{\sigma_n}{\sqrt{M}} \tag{3.18}$$

This function is plotted for various lengths of n, which when blocks are long enough to be independent of each other should yield a plateau in BSE, giving the "true standard error" of the function at hand.

Other notes on convergence

IDPs are known to be flexible proteins, why fluctuations on structural properties should be expected, giving large standard deviations. A timesseries of a (structural) property may not therefore show a singular value to be attained over time, as one could expect from a globular protein, but one should look at the major trends.

Ideally, a simulation is run for a long period of time, so that correct sampling is achieved irregardless of starting position. In practice, this may not always be true, due to aforementioned reasons. One should therefore start several simulations of the same system with slightly different starting positions (these different simulations of the same system is denoted "replicates"). In this Thesis, this is not strictly done - the same starting configuration is used, but the starting velocities are different, why the different replicates will, at least initially, evolve into slightly different structures. Comparing different replicates can also be a way to judge whether the simulation has reached convergence.

Despite the above tools available for deciding whether a simulation has been run for long enough, there is no "one-size fits all", simple visual inspection may sometimes be sufficient to determine if a simulation observable is systematically changing and one should consider what is "good enough", why communicating how analysis was performed and how interpretations are made is of high importance. [109]

4.4 Principal components analysis

Principal components analysis (PCA) is a general method for data analysis. It can be used to get an overview of a very large data set and find patterns in the data. A simulation trajectory can indeed contain large amounts of data: 3N - 6 data points for each snapshot

in a system with N particles (simulations with N being on the order 10^5-10^6 is today not uncommon, in this Thesis, a system with N $\approx$ 600 000 will be presented). PCA takes a data set, here represented by a matrix denoted X, consisting of a number of observations or measurements (corresponding to the number of rows of the matrix X) and a number of variables or features (corresponding to the columns of the matrix X), and transforms the data set into two matrices X = T.P, where T (called the scores, describes the observations) has the same number of rows as the number of observations and A columns, while P (called the loadings, describes the variables) has A rows and the same number of columns as the number of variables. A is the number of principal components (PCs). The beauty of this transformation is that one can use fewer PCs than the original number of variables, as the PCs are constructed in such a way that each successive PC describes as much variance or information content as possible. This is measured by the eigenvalue of each PC, which divided by the sum of squares of the data matrix X yields the proportion of variation covered by the PC. The reduction in the number of variables (or dimensions) can be large - in this Thesis, a PCA of a data set consisting of 48 spectra (being the observations), each having more than 1000 q-values (being the variables/features in this case) could be reduced to two PCs, easily shown in a 2D-plot, while retaining 99.5 % of the variance of the data set. This allows an overview of how different observations relate to each other. The PCs themselves are abstract mathematical entities, being linear combinations of the original variables and are orthogonal to each other.

Energy surface

The PCA of a trajectory can show how different structures in a simulation relate to each other, however, with some additional processing, one can also find which classes of structures that may be more important and identify barriers between different conformational classes. This is achieved by, for each structure, computing a conditional free energy,

$$E(\mathbf{r}) = -RTln\frac{P(\mathbf{r})}{P_{max}} \tag{3.19}$$

where $E(\mathbf{r})$ is the conditional free energy of structure with coordinates $\mathbf{r}$, R is the gas constant, T is the temperature, $P(\mathbf{r})$ is the probability density function and P_{max} is the maximum value of the probabiliy density function. Assigning the conditional free energy to the structures in the PCA plot as a heat-map yields a so-called energy surface, or energy landscape. In this Thesis, the method of Campos and Baptista [110] was used to compute the energy surface. The PCA is performed not on the whole set of coordinates, but on the protein backbone atom coordinates, after translational and rotational least-squares fitting to a central structures. This central structure (which should not be confused with some

kind of "equilibrium structure" often found for globular proteins) is found by computing a dispersion measure D_i, defined for structure i as

$$D_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} rmsd_{ij}^2 \tag{3.20}$$

where n is the number of structures considered and *rmsd* is the root-mean squared distance between the backbone atoms of structures i and j. The structure with lowest dispersion measure, i.e. the one structure most similar to other structures, is chosen as the central structure. The probability density function is estimated using a Gaussian kernel estimator. A possible pitfall when using this method to analyze a simulation is whether the two first PCs cover only a part of all the variation of the data set. The identified conformational classes from the two first PC may not then be representative of the whole data set.

4.5 Ramachandran plot

There are other ways to visualize and compare structural ensembles with each other. One is the Ramachandran plot, [111] which is based on the dihedral angles ϕ, ψ plotted against each other. It has been shown that certain regions of dihedral angle values corresponded to secondary structure values, as well as some regions being "forbidden", due to steric hindrance. One definition of how dihedral angles and secondary structure categories connect can be found in Hollingsworth and Karplus [112], which put α-helices to be found near the point ϕ, ψ = (-63, -43), with most variations found within $\pm$ 15° of this peak, the 3_{10} helix in the point ϕ, ψ = (-60, -25), the β-strand in the point ϕ, ψ = (-120, +130) while spanning a range of 80° in both angles, and the poly-proline II structure centered in ϕ, ψ = (-65, +145) with a range of 50° in both angles. It should however be stressed that categories of secondary structure and their position in the Ramachandran plot may vary in the literature.

4.6 Brief note on comparative performance between MC and MD

There are some studies that have compared using the MC methodology with the MD methodology, as to discuss which scheme is the best choice. A study of liquid hexane found MC to be superior, with MC being 2-3 more efficient than MD. [113] Another study, where a protein was simulated, found MD to have 1.5 larger sampling efficiency. [114] For general arguments, it is claimed that MD better accounts for collective motions, as found in longer chains, while MC can more easily transition across large energy barriers. Thus, which method is better can be considered system dependent. However, it might be more

important to consider the implementation of each method, as noted by Jorgensen and Tirado-Rives [113] when considering an older comparison between MC and MD. [115]

5 Coarse-graining

As stated earlier, the computational complexity is, without special techniques, $O(N^2)$. Molecular systems can be quite large, why some systems may be out of computational range. A technique to still be able to model such systems is the use of coarse-graining. Here, several atoms are considered to be part of one larger pseudo-atom, or bead. Interactions of the individual atoms are disregarded, and an effective interaction of all atoms constituting the pseudo-atom is used instead. This reduces the number of particles in the simulation, at the cost of molecular detail. An example of coarse-graining is shown in Fig. 3.1.

The method of coarse-graining has been around since the 1970s, with a prominent example being a coarse-grained model published in 1975 by Levitt and Warshel. [116]

Not only does the computational feasibility increase by the decreasing of the number of particles, but the coarse-graining also smoothens the free energy surface. [117] This decreases the risk of the simulation being trapped in local energy minima, and shortens the simulation time necessary to reach convergence. [118] The smoother free energy surface also makes it possible to have a larger integration steps. All these factors make it possible to speed up a simulation by 2-5 orders of magnitude. [119] Several different levels of coarse-graining can be considered. Models mapping 4 heavy atoms to one bead, [120–122] or a whole amino acid to one bead, [23, 123] or two beads, [124] exist, though it is not a requirement that the mapping to be strict. For example, there exists hybrid-approaches where part of the protein has an atomistic representation, and other parts has a coarse-grained representation. [125] The model of choice depends on the property of interest and available computer power.

This Thesis considers to a great extent a coarse-grained model with a one-amino-acid, one-bead mapping. This excludes, for example, the study of secondary structure, which requires a fairly high level of description detail for the protein back-bone atoms.

A general problem of coarse-grained models is, apart from the inability of computing some properties (depending on level of coarse-graining), the transferability of the models. [126] Care is therefore needed in the choice of model, which depends on the problem at hand.

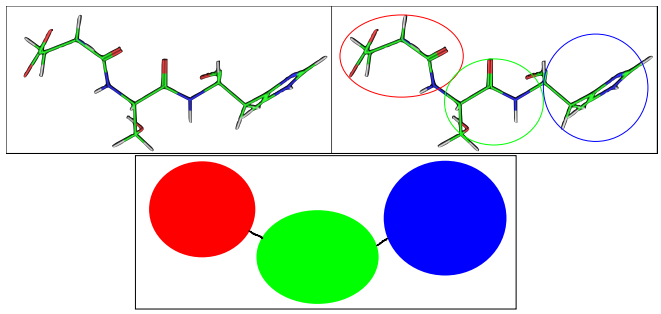


Figure 3.1: Example of coarse-graining. In the top-left figure, an atomistic representation of a molecule (three amino acids forming a peptide) is used. In the top-right figure, we form pseudo-particles out of the individual atoms, in this case a one-bead-per-amino acid mapping. Lastly, in the bottom figure, we have achieved a coarse-grained representation of the original molecule. No details of side-chains or any other internal degrees of freedom are presented, but a significant reduction in the number of particles has been achieved, reducing the computational cost of modelling the molecule.

6 Computation of observables

After performing a simulation, one computes the properties of interests. This section gives a brief overview of how this computation is done for a few select properties. In general, one usually need a so-called forward function to connect the simulation-produced ensemble of structures (trajectory) with what is observed in experiment. These may sometimes need parameterization and/or be approximate in nature, why this can sometimes be a source of error when comparing simulation with experiment.

6.1 SAXS observables

From SAXS experiments, one can obtain the radius of gyration, which can be computed from a simulation according to Eq. 2.13. However, one can also compute the whole scattering curve. In general (applying to both SAXS and neutron scattering), for a particular scattering orientation,

$$I(q) = |\sum_{j}^{N} b_j exp(-iqr_j)|^2 \tag{3.21}$$

with b_j being the generalized scattering length of particle j, N the total number of particles and r_j being the position of particle j. For SAXS, b_j implicitly depends on q. An integration over all orientations (with the assumption that there is no preference in the molecular orientation) yields the Debye formula for spherical scatterers:

$$S(q) = \sum_{j=1}^{N_j} \sum_{i=1}^{N_i} f_i(q) f_j(q) \frac{sinqr_{ij}}{qr_{ij}} \tag{3.22}$$

where f is the form factor for particle i, j, and r_{ij} is the distance between them. This formula is implemented in, for example, the software FoXS. [127, 128] Different software for the calculation of SAXS profiles may rely on different approximations on the Debye formula, or considers hydration in different manners. [129] It should also be noted that even when using the same software, but changing input parameters for the hydration shell may cause different results, on the order 5-10 %. [130] For the bead-necklace model used in this Thesis, all particles are assumed to be identical scattering objects, for which the scattering curve is computed according to

$$S(q) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} exp(i\mathbf{q} \cdot \mathbf{r}_j) \right|^2 \right\rangle, \tag{3.23}$$

which can be computed as a sum of partial structure factors S_{ij},

$$S(q) = \sum_{j=1}^{N_j} \sum_{i=1}^{N_i} \frac{(N_i N_j)^{1/2}}{N} S_{ij}(q), \tag{3.24}$$

which are in turn computed via

$$S_{ij}(q) = \left\langle \frac{1}{(N_i N_j)^{1/2}} \left[\sum_{i=1}^{N_i} exp(i\mathbf{q} \cdot \mathbf{r}_i) \right] \left[\sum_{j=1}^{N_j} exp(-i\mathbf{q} \cdot \mathbf{r}_j) \right] \right\rangle. \tag{3.25}$$

For this structure factor to be comparable with experiment, a scaling factor is needed, which is not computed, but the curve is fitted to the experimental data, with the scaling factor as the only fitting parameter.

6.2 Diffusion

Diffusion can be computed in several ways from simulation. One alternative is time-integrating the velocity autocorrelation function, however, in this Thesis, the Einstein relation

$$D = \lim_{t\to\infty} \frac{d}{dt}\frac{1}{6}\langle |r_i(t) - r_i(0)|^2 \rangle \tag{3.26}$$

has been used, where the $\langle ... \rangle$ expression is the mean square displacement (MSD). The relation is valid at long times, and requires the MSD vs. t plot to be linear. Therefore, one selects the region in the MSD vs t that displays linear behaviour. Deviations from linearity can, for example, be a consequence of poor sampling (mainly for large t) and non-diffusive behaviour. It has been stated that there exists no objective method to determine this region. [131] In this work, linearity of the chosen region was confirmed by computing R^2. Further, diffusion is known to be affected by finite box-size effects. [132] A way to adjust for this is via Eq. 3.27:

$$D_0 = D_{PBC} + \frac{k_B T \xi}{6\pi\nu\, L} \tag{3.27}$$

where D_0 is the true/corrected diffusion, D_{PBC} is the "raw" diffusion obtained from simulation, k_B is the Boltzmann constant, ξ is a constant dependent on the box geometry, ν is the viscosity of the solvent and L is the length of the side of the unit cell.

Another way to compute diffusion from simulation snapshots is through the HYDROPRO algorithm by Ortega *et al.* [133] This algorithm replaces the surface of particles with a number of small beads, which are used to compute properties through the primary hydrodynamic model, using frictional coefficients for the beads. [134]

6.3 Circular Dichroism

Direct computation of CD spectra from simulation

There are several competing suggestions to compute a circular dichroism spectra from a simulation. One of the more recent is a data-driven approach (called "SESCA"), using a set of reference proteins where both structure (from the Protein Data Bank) and CD spectra are known. [135] These were used to create a collection of "basis sets", from which spectra are computed according to

$$S_j^{calc}(\lambda) = \sum_{k=1}^{K}\sum_{i=1}^{F} W_{jk}\alpha_{ki}B_i(\lambda) \tag{3.28}$$

where S_j^{calc} is the spectra to be calculated, λ is the wavelength, K is the number of secondary structure elements, F is the number of basis spectra, W_{jk} is the fraction of residues in protein

j classified as the category of secondary structure k, α_{ki} is an assignment factor, giving the factor for the contribution of secondary structure element k from basis spectra i. α_{ki} was the main optimization target in this scheme. The above computation only considered the backbone-structure to be used for computing CD spectra, however, additional basis sets were also constructed to include the contribution of protein side chains to the CD spectra, though this procedure was performed independently of the rest of the protein.

Indirect comparison

Rather than computing a CD spectrum from a simulation, one can instead analyse the experimental spectrum to estimate the relative amounts of secondary structure (as noted in Chapter 4). The secondary structure from simulation is then computed, according to a chosen definition of secondary structure, and the two amounts are compared with each other. For different choices of defintions of secondary structures, DSSP [136] is a common choice which is based on hydrogen bonding, while an alternative called DISICL is based on dihedral angles only. [137]

Chapter 4

Experimental background

1 Sample preparation

A possible discrepancy between experimental measurements and simulated predictions is a difference between the conditions in either case. This could relate to a lack of control in the experimental case, for example there being unknown impurities in the samples (which may also be important for reproducible results for a purebred experimentalist). Care must therefore be taken when preparing samples. It should however be pointed out that 100% purity may not be practically achieveable.

Protein samples used in this Thesis were mainly bought from companies using the solid-phase synthesis method. Samples were delivered as a lyophilized powder, which however contained TFA (CF_3COONa) and other monovalent salts (e.g. NaCl, KCl), according to the manufacturer. To achieve pure samples, these samples were dialyzed against Milli-Q water using 500-1000 Da membranes, the ratio of Milli-Q/sample solution volume being between 250 - 500. The Milli-Q water was changed to fresh Milli-Q water 4-5 times during the course of two days. After dialysis, the sample solution was freeze-dried, and then ready for use in a solution.

The buffer solution used was mainly 20 mM Tris, together with NaCl salt at pH=7. An exception is the circular dichroism measurements, where NaF salt was used instead, due to the high absorbance of Cl in circular dichroism spectroscopy. As well, before measurements with circular dichroism, the sample solution was filtered (this also applied for dynamic light spectroscopy measurements).

2 Small angle X-ray scattering

Small angle X-ray scattering (SAXS) is the use of X-rays (light with a short wavelength <0.3 nm) to irradiate a sample and infer particle structure based on the angle-dependent distribution of the scattered X-rays at small angles. This is demonstrated in Fig. 4.1, where a crude, constructed pattern of scattering is used to demonstrate how the spectrum is produced from the pattern of scattering.

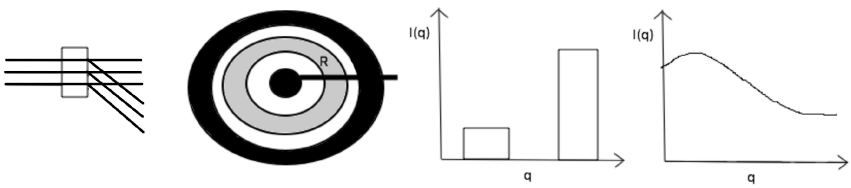


Figure 4.1: *Far left:* When X-rays irradiate the sample, they may interact with the sample, scattering in a different angle compared to the incident angle. *Mid left:* At the detector, the intensity at different angles (the distance from the center in any direction) is recorded. *Mid right:* The intensity found at the detector is radially averaged and displayed as a function of q, a wavelength-independent metric of angle. This example shows how the mid-left figure would approximately be displayed. *Far right:* A more realistic example of a scattering curve.

An incident X-ray may scatter if there is an interaction with the sample, otherwise it may just go straight through the sample. Mainly, the interaction that produces scattering is caused by collisions with electrons in the sample. It is possible that an X-ray scatters more than once (i.e. interacting with more than one particle in the sample), but this is usually neglected.

If the particles are randomly oriented in the sample, the intensity found at the detector will be the same in all directions. The X-ray scattering can be measured as a function of the angle θ (which at the detector will correspond to a distance R from the mid-point, but which depends on the length between sample and detector), but to relate to the fact that the angle may depend on the wavelength λ of the X-rays used, a measure q is used:

$$q = \frac{4\pi}{\lambda} sin(\theta) \tag{4.1}$$

which is denoted ”momentum transfer” or ”length of scattering vector”. This vector can be related to the distance probed in the sample, via

$$d = \frac{2\pi}{q} \tag{4.2}$$

where d is distance. For highly ordered and peridiodic systems, pronounced peaks can be visible in the spectra, which are referred to as Bragg peaks. Eq. 4.2 is derived from Bragg's law, where aligned particles are considered.

The intensity at any q for a given sample in a solution (or matrix) is dependent several factors, as seen below:

$$\Delta I(q) = N I_0 (\Delta p)^2 V^2 P(q) S(q) \tag{4.3}$$

where N is the number of particles, I_0 in the intensity of the incoming beam, Δp is the difference in electron density between the sample and the solution (or matrix), V is the volume of a particle, P(q) is the form factor and S(q) is the structure factor. Eq. 4.3 holds for monodisperse samples; for polydisperse samples, a summation of all particle dependent terms is required instead of a multiplication with N.

A few observations should be noted from this equation. If N is small (i.e. there is a low sample concentration) the scattered intensity will be low. To achieve spectra with reasonable resolution, instruments with a large incoming intensity (I_0) may thus be necessary for dilute samples, such as those instruments found at synchrotrons. As well, low intensity is generated if the difference in electron density between sample and solution (matrix) is low, making some samples hard to distinguish. The intensity increases on the square with the volume of the sample particle, and as the volume of a (spherical) sample increase with the cube of the radius, large sample particles will dominate with a signal to the sixth exponent, therefore even small amounts of aggregates in protein solutions can completely dominate the spectrum. This can be considered both an advantage or a disadvantage: smaller aggregates are fairly easy to discover, while at the same time, signal from the larger volume of the sample is obscured.

The form factor P(q) describes the shape of the particles, while the structure factor is a measure of relative distance between particles in a solution, with shorter distances yielding larger structure factor. The relative distance between particles will depend on the concentration of particles in the solution and the interactions of the particles. For repulsive particles, the relative distance between particles will be larger than with non-interacting particles, resulting in smaller structure factor. Likewise, with attractive interactions between particles, relative distances will be shorter, yielding larger structure factor. For very dilute systems, where particles can be considered far away and interactions small, the structure factor is approximately 1, so very dilute systems will yield the form factor, with all other factors collected into a constant. The scattering power of a sample, referred to as scattering length density (SLD), is related to the electron density of the sample and the energy of the X-rays. The energy dependence becomes important when the energy is close to the absorption edge of a sample, which is utilized in so-called Anomalous Scattering Angle X-ray Scattering

(ASAXS), which is not considered in this thesis. Biological samples have usually a weak anomalous effect.

X-rays can be scattered in a coherent manner or in a incoherent manner. Coherent scattering is when the scattered X-rays (or any other type of scattering) have a certain degree of regularity, or rather, the waves scattered from different atoms interact with each other, making the final "outbound" scattering dependent on the relative distance between the atoms, encoding information about the structure in a sample. Incoherent scattering on the other hand has the waves scattered being independent from each other, so that waves scattered from different atoms do not interact with each other.

2.1 Guinier analysis

The form factor can, at very small angles, be considered Gaussian. In the Guinier analysis [138], this Gaussian is approximated by

$$P(q) \approx a_o e^{-\frac{R_g^2 q^2}{3}} \tag{4.4}$$

where a_0 is the extrapolated intensity at zero angle ($I(0)$), when considering the actual intensity rather than P(q). Importantly, this expression can be linearized so the parameters are supplied by linear fitting. Thus, the radius of gyration can be obtained through the use of SAXS, if the sample is dilute so that there is no contribution from the structure factor.

Exactly how small angle (q) that is necessary depends on the sample studied, but for well-folded proteins, a rule of $qR_g < 1.3$ is used, while for IDPs, a rule of $qR_g < 0.8$ is used. [139]

There exists suggestions for an "extended" Guinier analysis. "Extended" in this context refers to the range of q-values that can be used in the analysis. Eq. 4.4 is expanded to be

$$P(q) \approx a_o e^{-\frac{R_g^2 q^2}{3} + \alpha q^4 R_g^4} \tag{4.5}$$

where α is a parameter, dependent on the system at hand. A parameterisation of this equation for IDPs, using simulations for parameterisation, yielded $\alpha = 0.0479(v - 0.212)$, where v is the polymer scaling exponent. [140]

The scattering intensity can also be related to the molecular weight of the particles, via the volume and the (electron) density of the particles. In theory, this can be calculated at any value of q according to

$$M_w = \frac{\Delta I(q)}{K_0(\Delta Z)^2 P(q) S(q)} \tag{4.6}$$

where K_0 collects constant terms and ΔZ is the electron density difference scaled with the density of the particle materials. With dilute systems, S(q)=1, and in order to avoid using a model of P(q), one can use the intensity at $q = 0$, where P(q) ($I(0)$) is found via the Guinier approximation.

2.2 Kratky plot

A Kratky plot is constructed by plotting the $(qR_g)^2 I(q)/I(0)$ vs qR_g [141]. With such a plot, the shape of the protein can be assessed, as the plot yields different behaviours for different (average) shape of the protein at high values of qR_g. For a globular protein, the plot is bell-shaped, while a gaussian chain exhibits a plateau with increasing qR_g, and a rod-like protein has an approximately linear behaviour with a positive slope. Examples of these are shown in Fig. 4.2.

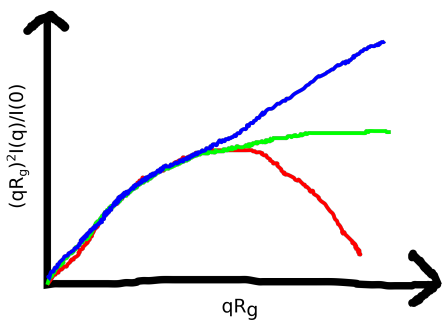


Figure 4.2: Sketch of examples of different shapes of proteins deduced from Kratky plot. Red curve: Globular protein. Green curve: Gaussian chain. Blue curve: Rod-like protein.

2.3 Fractal dimensionality number

The fractal dimensionality number [142] considers, just as the Kratky plot, the overall shape of a particle. Fractal dimensionality refers to the number of dimensions a particle occupies; 1 is a straight line, 2 is a plane and 3 is a sphere, though, as the name implies, fractal numbers are allowed. The number is computed by estimating the slope of log(I) vs. log(q) at higher values of q.

3 Quasi elastic neutron scattering

Quasi elastic neutron scattering (QENS) differs from (small angle) X-ray scattering in two ways: X-rays are changed to neutrons and the change in energy in the incident and scattered neutrons is measured. The scattering power of neutrons does not depend on electron density as in X-ray scattering, rather, it depends on the neutron interaction with the nucleus of the sample. This gives neutron scattering two particular features: It is sensitive to isotopes, and the scattering power is irregular with increasing atomic number of the elements. Scattering lengths for select elements are found in Table 4.1

Table 4.1: Neutron scattering lengths for a few select elements. Data from Bee (1988).[143]

Element	Coherent scattering length (10^{-15} m)	Incoherent scattering length (10^{-15} m)
H	-3.7423	25.217
D	6.674	4.033
C	6.6535	0
O	5.805	0
N	9.37	1.98

In QENS, it is mainly the incoherent scattering that is measured. As an example, for Histatin 5, with a chemical formula of $C_{133}H_{195}N_{51}O_{33}$, $\approx$ 86 % of the scattering would be incoherent. For a protein in water solution, the water can be exchanged to deuterium oxide to decrease the signal from the solvent.

The scattering function for incoherent scattering is defined as

$$S_{inc}(q,w) = \frac{1}{2\pi N}\int_{-\infty}^{+\infty} dt exp(-i\omega t)\sum_{i=1}^{N}\left\langle exp[iq(R_i(t) - R_i(0))]\right\rangle \tag{4.7}$$

where ω is the energy (frequency), N is the number of particles and $R_i(t)$ is the position of particle i at time t. This expression can be treated with an inverse Fourier transform twice to reveal the van Hove self-correlation function,

$$G(r,t) = \frac{1}{N}\int\left\langle\sum_i \delta(r - R_i(0))\delta(r - R_i(t))\right\rangle dr \tag{4.8}$$

which reveals that QENS report on the autocorrelation of position, i.e. dynamics.

Assuming a diffusive process, going from a van Hove function back to scattering through treatment Fourier transform twice, one arrives at a Lorentzian function,

$$S(q,w) = \frac{1}{\pi}\frac{\gamma}{\gamma^2+\omega^2} \tag{4.9}$$

where γ is the half-width at half maximum (HWHM). γ can be related to diffusion in different ways depending on the diffusive process. For a freely diffusing particle (Fickian diffusion), γ = Dq^2. For a jump-diffusion process, one could use the model of Singwi and Sjölander, [144]

$$\Gamma_{jump}(q) = \frac{D_{diff}q^2}{1+D_{diff}\tau q^2} \tag{4.10}$$

where D_{diff} is the diffusion coefficient and τ is the time between jumps. A QENS experiments probes dynamics during a specific time window (through the Fourier transform of energy into time - being very strict, it does not directly probe dynamics). This time window is determined by the energy range of the instrument. This is in contrast to other methods, which might only measure long-term diffusion. Thus, the diffusion coefficient obtained from QENS may not be directly comparable with that obtained from other techniques.

Several different diffusive processes may take place in a sample. Modelling-wise, one can add and/or convolute together several Lorentzian functions to account for each process. The process can occur to different extents, why a factor is used to account for the relative contribution of each diffusive process. This factor is referred to as Elastic Incoherent Structure Factor (EISF), also often denoted A_0. While the incoherent scattering spectra is insensitive to the value of q, the EISF is not. The EISF can therefore give information on the geometric confinement of motions.

4 Dynamic light scattering

In SAXS and QENS, the scattering (as found by the detector) is measured in a time-independent manner. In dynamic light scattering (DLS), the time-dependency is taken into account, measuring intensity (for a given q-value) as a function of time. From this data, one can construct an intensity correlation function,

$$g_2(\tau) = \frac{\left\langle I(t)I(t+\tau)\right\rangle}{\left\langle I(t)\right\rangle^2} \tag{4.11}$$

where I is the intensity, t is any given time, τ is the lag time and the denominator is a normalisation factor. This correlation function is related to a an electric field correlation function, which correlates the movements of particles relative to each other. Considering the normalized electric field correlation function, denoted g_1, the relation is

$$g_2(\tau) = \mathrm{B} + \beta|g_1(\tau)|^2 \tag{4.12}$$

where B is a baseline, and β is a constant depending on the scattering particles and the instrumentation. For Brownian motion of monodisperse particles, the electric field correlation function is an exponentially decaying function, $e^{-\Gamma\tau}$, where Γ is the decay constant. This constant is related to the diffusion of the particle by $\Gamma = -Dq^2$. The hydrodynamic radius can thereafter be computed from the diffusion by the Stokes-Einstein equation,

$$R_h = \frac{k_\mathrm{B} T}{6\pi\eta D} \tag{4.13}$$

where k_B is the Boltzmann constant, T is the temperature and η is the viscosity of the solution. For a polydisperse system, a distribution of decaying exponentials is needed to represent the electric field function. This can be formulated as a sum of decay exponentials,

$$g_1(\tau) = \sum_{i=1}^{M} a_i exp(-\Gamma_i \tau) \tag{4.14}$$

where M is the number of exponentials to be used and a_i describes the distribution of the exponentials.

The contribution to the intensity of different particles depends on the size of the particles. For a solution consisting of two particles with size a and b, the relative intensity from particle a would be

$$\%I_a = \frac{a^6 N_a \cdot 100}{a^6 N_a + b^6 N_b} \tag{4.15}$$

where N_a, N_b is the number of particles a and b, respectively. The scaling of size to the power of six means that larger particles may dominate the scattered intensity, even if the number of large particles is small.

5 Circular dichroism

Light can be viewed as a wave, which while propagating forward oscillates in the planes perpendicular to the direction of propagation. [145] The oscilliation is usually isotropic, i.e. the oscilliations are in all possible directions. However, when this does not apply, the light is said to be polarised. The polarisation can be linear, so that the light only oscillate in one direction, or the polarisation can be circular, so the direction rotates about the direction of propagation. Circular polarisation can be clockwise (right handed, R) or counter-clockwise (left handed, L). Circular dichoism (CD) is the differential absorption of these when light passes through a sample. A sample that can absorb light (contains a chromophore) needs to be chiral in order to produce a difference in absorption between the R and L polarised light. It is this difference ($\Delta A = A_L - A_R$) that is monitored for different wavelengths in the CD experiment.

If one would look in the propagating direction, a polarised light would oscillate in an elliptical manner, as L and R have different magnitudes. A common unit used for the polarisation is ellipticity (θ), which can be computed as $\theta = tan^{-1} b/a$, where b and a are the minor and major axes of the ellipse. This is related to the absorbance ΔA through $\theta = 32.98\Delta A$. [146] Further, the concentration of sample and, in the case of proteins, the number of peptide bonds may affect signal strength, why the measure $[\theta_{MRW}]$ is used:

$$[\theta_{MRW}] = \frac{MRW \times \theta}{l \times c} \tag{4.16}$$

where MRW is the Mean Residue Weight in Da, l is the pathlength in mm and c is the concentration in g/ml.

In proteins, it is the peptide bond (mainly absorbs in the 170-240 nm region), aromatic amino acid side chains and disulphide bonds (mainly absorbs in the 250-320 nm region) that are considered in CD. In particular, the peptide bond, gives information about the secondary structure in a protein. For example, estimates for alpha-helical content can be obtained by using the values at wavelengths 208 nm and 222 nm, and CD-spectra associated with different types of secondary structure are available. [146] A few examples of spectra of proteins with a high degree of specific secondary structures are found in Figure 4.3.

There are also algorithms available that uses a range of wavelengths in the CD-spectra to infer estimates of several categories of secondary structure. [151]

Even if IDPs lack distinct tertiary structure, there can still be transient secondary structure, why CD is a useful tool in the case of IDPs. However, CD only yields averages of the secondary structure.

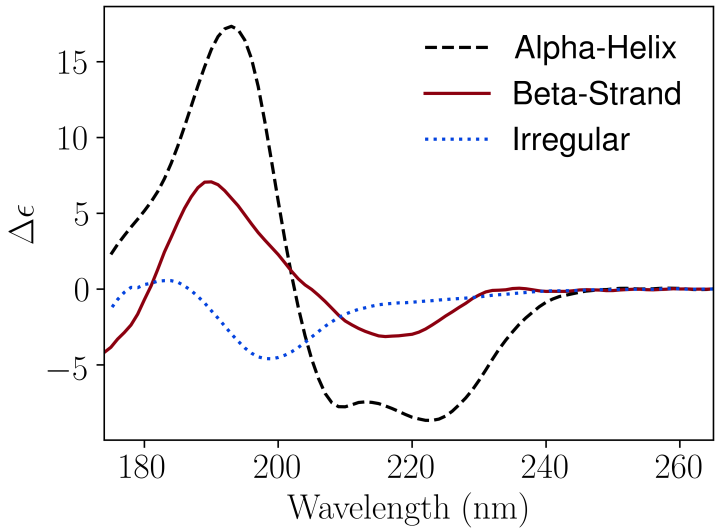


Figure 4.3: Example CD spectra using proteins acquired from the PCDDB resource, [147], each protein representing a secondary structure category. The alpha-helix structure is represented by myoglobin [148] (DSSP-value of 0.742 in the α-helix category), beta-strand structure is represented by OmpG [149] (DSSP-value of 0.676 in the β-strand category), and irregular/loop structure is represented by Translocated Actin Recruiting Phosphoprotein [150] (DSSP-value of 0.71 in the irregular/loop category).

An important issue to consider when using CD is the use of buffers. Some buffers absorb in a wavelength-dependent manner, causing signal (and signal-to-noise ratio) of the species of interest to decrease. The same problem may arise if sample concentration is high, leading to high absorbance and low signal-to-noise ratios, or if the sample is complex in the way of containing several species, where some are particularly absorbing.

Chapter 5

The Research

Proteins are usually investigated at dilute concentrations, but in biological settings, macromolecular concentrations are high, in the range of 300-400 mg/ml. [32] The high concentration may affect both the structure and the dynamics of proteins. In the case of IDPs, having no distinct 3D-structure in dilute conditions, different outcomes can be realized due to the effect of crowding encountered in high macromolecular concentrations.

In this Thesis, the crowding effect has been investigated for two IDPs: The saliva protein Hst5 (main focus) and the artificial dimer of Hst5. Where Hst5 is found *in vivo* with disease-prevention functions, the dimer was chosen as a way to investigate if the crowding effect could be dependent on the sequence-length of the protein. Hst5 has in earlier research been investigated by several methods in dilute conditions, while the dimer is a novel protein.

The approach used here was a combined experimental and computational one. With disordered proteins, some experimental techniques are not suitable to use, such as crystallographic techniques. Instead, SAXS have been the main method of choice, being able to account for the shape and size of proteins, with the downside of only providing average structures. Computer modelling is used to achieve more detailed information than available from experimental data. With high protein concentration follows high computational power costs. This is countered by employing coarse-grained models. A potential pitfall of using modelling techniques is that models may not be well-parametrized or suitable for the model systems considered. For example, many atomistic models have previously been found to be inadequate for modelling IDPs, [80] though improvements have been made in later years. [152] The computer models thus need to be confirmed to agree with experimental data, before conclusions are drawn from the results of the simulations.

1 SAXS

For both IDPs investigated, a protein concentration series was measured with SAXS. Temperature and salt concentration was as well varied, to assess the influence of these parameters as protein concentration increases. In both cases, in the high salt concentration domain, there was a range of protein concentration where the proteins remained monodisperse. In the case of Hst5, this was up to 50 mg/ml, while for the dimer, it was found to be about 25 mg/ml, though a vague hint of aggregation was noted already at this concentration. At higher protein concentration, both aggregation and interparticle effects were found. This change in point of aggregation for different protein lengths is in qualitative agreement with concepts from polymer physics. [153]

At lower salt concentration, interparticle effects were visible at very low protein concentration, if not the lowest measured, for both IDPs. The interparticle effects were mainly repulsive effects, being stronger at lower salt conditions at all protein concentrations. As salt screens electrostatic effects, the repulsion is considered an electrostatic effect.

No effect was found when varying temperature in the case of Hst5, while a modest effect was found in the case of the dimer. Temperature had, for the dimer, a greater effect in the case of high protein concentration, which might be connected to the presence of aggregates at those concentrations.

A cell is a heterogenous environment, where not only the protein itself may contribute to make the environment crowded. As a model for other kinds of macromolecules with little "direct" interactions (i.e. being inert) with proteins, poly-ethylene glycol (PEG) at various sizes and Ficoll70® was used at increasing concentrations in solutions together with a constant concentration of Hst5. These investigations showed that, up to a crowder concentration of 100 mg/ml, there was no discernable impact on Hst5. The crowders themselves experienced repulsive interactions, but Hst5 was unaffected, at least for the PEG crowders. Ficoll70® had such a large scattering intensity that the Hst5 signal was partly "drowned" in the noise, why the effect of Ficoll70® as a crowder is somewhat uncertain from the SAXS experiments.

2 Diffusive properties

Concentrated solutions of Hst5 were measured with quasi-elastic neutron scattering (QENS), giving information about dynamical properties. QENS uses neutrons, which are fairly penetrative, so in order to get sufficient scattering signal, a fairly high concentration is needed to get a good enough signal. The lowest concentration of Hst5 considered was 50 mg/ml in this case. From the QENS measurements, it was found that with increasing Hst5 con-

centration, the diffusion decreased significantly - at 200 mg/ml protein concentration, the diffusion was less than 40 % of the diffusion found at 50 mg/ml. In part, this can be explained with the aggregation found by SAXS. The QENS measurements also encompassed different temperatures and salt concentrations. The temperature dependence of the diffusion was found to trivially follow Stokes-Einstein behaviour. Increasing salt concentration yielded slower diffusion, which could not be referred to changing solvent properties alone. A possible explaination to the salt dependence is to look at the impact on salt on structure (as observed from simulations), which show slightly larger R_g at lower salt concentrations, which speculatively would be equivalent to more elongated structures, known to have faster diffusion. At low protein concentration (10 mg/ml), Hst5 has also been investigated with dynamic light scattering (DLS). The value of diffusion obtained, 18.9 $Å^2$/ns, is larger than the value obtained from QENS at 50 mg/ml (16.8 $Å^2$/ns). Given the trend of decreasing diffusion with increasing protein concentration, as observed from QENS, this was an expected result. However, even if the concentration would have been the same in both instances, differences may have appeared due to the differences between the techniques, which inherently measure slightly different diffusional modes. The measurement of Hst5 with DLS also revealed two relaxation processes, corresponding to one smaller and one larger particle in the solution. Further investigation indicated that there exist, to a small extent, dynamical aggregates in the solution. This had not been previously detected by SAXS, and is largely ignored, due to the small population found. Using DLS, the impact of PEG, Ficoll® crowding on Hst5 was investigated. A complication here is that DLS mainly captures the behaviour of larger particles. With the exception of the smallest PEG specie investigated, PEG2K, all crowders were larger than Hst5. Thus, the measurements would rather display the impact of Hst5 on the diffusional properties of the crowders. For all crowding agents, no effect of adding Hst5 to a crowder solution at low concentration was found, in line with the SAXS results. At higher crowder concentrations, there was a hint of an effect using PEG4K and PEG6K, but nothing definitive. Ficoll® on the other hand showed a clear effect, with the diffusion decreasing, compared with the "pure" crowder solution. For the pure crowder solution, diffusion increased with increasing crowder concentration. This would be indicative of a polymer network/entanglement, as can be formed in the semi-dilute regime. Given the relative change upon the addition of Hst5, the hypothesis is that Hst5 modulates the polymer network, though without completely breaking it up, given that even with the added Hst5, diffusion still increases with increasing crowder concentration.

3 Bead-necklace model

Whenever a SAXS measurement was performed, a corresponding simulation was performed using a bead-necklace model that has previously been successful for predicting the structure

of several IDPs (including Hst5), though at dilute conditions. [154] For Hst5, as protein concentration increased, the experiment/model agreement decreased, up to the point where aggregation was experimentally found. At this point, agreement increased momentarily, but further increases in protein concentration worsened agreement. The discrepancy was mainly in regards to repulsive interactions, where electrostatic effects were excessive at lower protein concentrations, but not solely at fault at higher protein interactions. The model predicted mainly a conserved structure at crowded conditions, and was considered the best coarse-grained model considered for Hst5 in regards of modelling crowded conditions.

Considering the dimer, the experiment/simulation agreement was worse than for Hst5 at lower protein concentration. However, as protein concentration increased, agreement became better, up to a point, thereafter agreement decreased with increasing protein concentration. Importantly, the radius of gyration was not accurately predicted. As well, the model broke down for low salt concentrations in the case of the dimer, in the sense that the system aggregated at lower protein concentrations. A hypothesis explaining this behaviour attributes the Tris buffer as contributing to the screening of interactions, just as the salt do, which several different buffers are known to do, to various extent.[155] This would affect the Debye-Hückel screening length. Additional simulations indicated aggregation to stop if the buffer was considered to contribute to the screening length to the same extent as if adding 5-10 mM of salt.

For simulations where Hst5 was crowded by the polymers PEG (of various lengths) and Ficoll70®, modelling the crowding agents as one sphere/crowder, there was only small changes with increasing crowder concentration. Agreement with SAXS also decreased as crowder concentration increased. Comparing with the self-crowded simulations, the relative change in R_g for Hst5 was larger when using spherical crowder, which is speculatively attributed to the unrealistic volume occupied by the crowding agents when using spherical crowders. Another difference between the self-crowded simulations and having spherical crowders was the way the models differed from experiment: The self-crowded model displayed, relative to experiment, repulsive interactions (lower intensity at low q, relative to experiment), while the simulations with spherical crowders showed attractive interactions (higher intensity at low q, relative to experiment). Using the bead-necklace model together with another bead-necklace model specifically developed for PEG chains by Xie *et al.*, [156] the crowding response was found to be milder, in terms of R_g, as the difference from the non-crowded value of 13.8 Å was smaller. However, comparing with SAXS curves, the model had mixed performance compared to the case of having spherical crowders.

Looking at a limited selection of seven IDPs, (and three phosphorylated IDPs), the bead-necklace model has been shown to have some degree of transferability. [154] Extending upon this result, an additional 22 IDPs were simulated, using a data set that had previously been used for benchmarking a model called SOP-IDP, with two beads per amino acid,

while building upon the bead-necklace model. [124] It was found that the bead-necklace model had similar performance as the latter model in terms of R_g, showing that a simple model can provide as good results as a more advanced model.

3.1 The MARTINI model

The 4-to-1 bead coarse-grained model of MARTINI was already at very low protein concentrations of Hst5 predicting both very compact structures and aggregation, in disagreement with experiment. Other studies have found similar behaviour, [157] and suggested improvements to the model. [158] Using the suggested improvements of Stark *et al.*, [158] better agreement on Hst5 dimensions (as found by SAXS) was observed. This model was further used for modelling the crowding response of Hst5 when subjected to PEG crowders. In line with experiment, the effect of crowding on Hst5 was negligible/small. Instead, compared to single-chain simulations, the PEG crowder was to a larger degree affected by the crowding. This would point to Hst5 being, relative to other polymers, resistant to crowding effects, at least in terms of structural changes. The MARTINI model with adjustments of Stark *et al.* was additionally benchmarked against the series of IDPs used for benchmarking the bead-necklace model and the SOP-IDP model. [124] This MARTINI model had less good predictability of the experimental data compared with the models benchmarked, which might be attributed to the MARTINI model being a more universal model, also being able to capture the behaviour of globular proteins at the cost of performance for the sub-class of disordered proteins.

4 Other models

The Ensemble Optimisation Method (EOM) [159] and the ProFasi model, [160] were also considered for the interpretation of SAXS data of Hst5 at crowded conditions. These however were not found to be suitable, and were not considered in the modelling of the dimer, due to the low performance of modelling Hst5.

EOM, which chooses protein structures out of a large ensemble to reproduce SAXS data, predicted the protein ensemble to be bimodal at low protein concentrations and was generally not handling high-salt conditions well. The poor suitability of EOM for Hst5 was later also confirmed by others. [161]

ProFasi, an implicit-solvent, atomistic model with a simplified interaction potential, showed good agreement with SAXS measurements of Hst5 at dilute conditions. However, it provided either too attractive or too weakly repulsive interactions at higher protein concentrations, which was considered a consequence of the model not considering long-range electrostat-

ics. This feature also only allowed comparisons with experimental data using high salt concentrations.

5 Secondary structure content

With the failure of the bead-necklace model in the case of the dimer, it was considered if there might be transient secondary structure present, which the bead-necklace model cannot account for, due to its coarse-grained nature. Circular dichroism was therefore performed, for both Hst5 and the dimer. The overall shape of the curves were similar, though the amplitude of peaks and troughs were larger in the case of the dimer. Using the BESTSEL algorithm [151] to determine the amount of secondary structure, the difference between the two IDPs were negligible, though the amount indicated was fairly high in absolute terms - 27 % was found to be anti-parallel β-structures.

6 Atomistic modelling

6.1 Hst5-dimer

With the failure of the bead-necklace model and the fairly high secondary structure content predicted by BESTSEL from CD-data, a fully atomistic force field with explicit solvent was used to produce a more accurate picture of the dimer, at single-chain conditions (corresponding to very dilute conditions), using the force field Amber99SBN-ILDN, [162] with a TIP4P-D water model. [100, 163] Previously, atomistic modelling of Hst5 (with the same force field) has been successful, therefore should, ostensibly, the modelling of the dimer have promise to be successful. Performing this modelling, the atomistic modelling yielded a similar size of the dimer as the bead-necklace model, and too attractive interactions comparing with the SAXS data.

Comparison with CD-data was somewhat inconclusive, as this comparison requires the use of either DSSP [136] + BESTSEL or the use of algorithms generating CD-spectra from simulations, which may have limited precision. In particular, different algorithms of the latter type were highly divergent. This shows the importance of further development of both CD-algorithms and atomistic force fields, and how a change of sequence length (even though the relative amino acid composition remains the same) can have a significant impact.

6.2 Crowded Hst5

Atomistic modelling of self-crowded Hst5 was performed using the Amberff99SB-*disp* with the accompanying TIP4P-derived water model and parameters for ions. [92] The main interest here was how diffusive properties would compare with the QENS data, but how structural properties would change with increasing protein concentration was also of interest, in particular since data was available for this kind of comparison from SAXS. In single-chain simulations, R_g was found to be 13.1 Å using this force field, fairly in line with previous studies which had obtained values of 12.9 and slightly below and above 12 Å (dependent on whether enhanced sampling was used or not). [164, 165] Increasing the protein concentration to 10 mg/ml yielded a R_g of 12.2 Å, and an increase to 50 mg/ml yielded a R_g of 12.9 Å. Given the differences in R_g reported from different studies, and the standard deviation being about 2 Å for the simulations in this Thesis, the data is interpreted as pointing to no difference in R_g due to crowding (up to the concentration considered). Additionally data supporting non-structural effects of the crowding was the visual observation of Ramachandran plots, indicative of secondary structure, not changing noticeably across the simulations of different protein concentrations. The translational diffusion coefficients were computed to be 19.8, 18.0 and 13.2 $Å^2$/ns for the single-chain, 10 mg/ml, and 50 mg/ml simulation, respectively. Specifically, the diffusion coefficients were computed to correspond to the coherence time of QENS-measurements, but additional considerations are necessary to compare with the QENS measurements, namely that the QENS measurements were conducted in deuterium, not in water, why a correction to this difference is necessary. This correction yielded translational diffusion coefficients of 16.0, 14.6, and 10.7 $Å^2$/ns, for single-chain, 10 mg/ml, and 50 mg/ml simulation, respectively. Comparing the simulations, there is a monotonic decrease in translational diffusion as protein concentration is increased, qualitatively in line with trends found by QENS measurements. This also indicates in particular that diffusion properties may be more sensitive to crowding than structural properties. However, for a comparison with QENS data, one also has to consider that QENS measures an apparent diffusion, where rotational and translational diffusion is convoluted. From another QENS study of an IDP, it was found that the ratio between translational diffusion and apparent diffusion was 1.27. [166] Applying this ratio for the 50 mg/ml simulation (the concentration at which experimental data is available), a value of 13.6 $Å^2$/ns was found, slower than the experimental value of 16.8 ± 0.7 $Å^2$/ns. It should however be emphasized that this comparison hinges on the ratio between translational diffusion and apparent diffusion found for another IDP being the same for Hst5. For reference, approximating Hst5 for an ellipsoid, the ratio is 1.7, which would put the diffusion on par with experiment, but the ellipsoid approximation has been found to be a poor model to predict diffusion properties of Hst5. Therefore, considering the approximate ratio for another IDP to be the better estimate, the force field is found to have slower dynamics than found by experiment.

7 Outlook

The research performed here show Hst5 to be, structurally, fairly resistant to crowding effects. One suggested reason for this resistance, the short length of Hst5, was considered by investigating the twice as long Hst5 dimer, which mainly showed a lower aggregation point, with respect to protein concentration compared to Hst5. However, this dimer is still a very small protein - indeed, the encyclopedia Britannica claims that the limit of what should be called a "protein" is at 50 amino acids, while anything smaller should be referred to as a "peptide", a category which even the dimer, at 48 amino acids, would belong to. [167] Thus, whether resistance to crowding is an effect of length is not yet settled. Many different models have been used in this Thesis. Often it has been found that these have semi-quantitative agreement with experiment, showing additional developments to be necessary but currently offering performance that can give initial insight into the nature of intrinsically disordered proteins.

8 References

[1] A. K. Dunker, M. M. Babu, E. Barbar, et al. What's in a name? why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*, 1(1):e24157, **2013**. doi: 10.4161/idp.24157. URL `https://doi.org/10.4161/idp.24157`. PMID: 28516007.

[2] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181 (4096):223–230, **1973**. doi: 10.1126/science.181.4096.223. URL `https://www.science.org/doi/abs/10.1126/science.181.4096.223`.

[3] V. N. Uversky and A. K. Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1804:1231–1264, **June 2010**.

[4] P. Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends in Biochemical Sciences*, 37(12):509 – 516, **2012**. ISSN 0968-0004. doi: https://doi.org/10.1016/j.tibs.2012.08.004. URL `http://www.sciencedirect.com/science/article/pii/S0968000412001259`.

[5] M. G. Iadanza, M. P. Jackson, E. W. Hewitt, et al. A new era for understanding amyloid structures and disease. *Nat. Rev. Mol. Cell Biol.*, page 1, **2018**.

[6] T. K. Karamanos, M. P. Jackson, A. N. Calabrese, et al. Structural mapping of oligomeric intermediates in an amyloid assembly pathway. *eLife*, 8:e46574, **2019**.

[7] P. Kulkarni and V. N. Uversky. Intrinsically disordered proteins and the janus challenge. *Biomolecules*, 8(179), **December 2018**.

[8] C. J. Oldfield, Y. Cheng, M. S. Cortese, et al. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6):1989–2000, **2005**.

[9] R. Pancsa and P. Tompa. Structural disorder in eukaryotes. *PLOS ONE*, 7(4):1–10, **04 2012**. doi: 10.1371/journal.pone.0034687. URL https://doi.org/10.1371/journal.pone.0034687.

[10] A. Campen, R. M. Williams, C. J. Brown, et al. Top-idp-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein and Peptide Letters*, 15(9):956–963, **2008**. ISSN 0929-8665. doi: doi:10.2174/092986608785849164. URL https://www.ingentaconnect.com/content/ben/ppl/2008/00000015/00000009/art00014.

[11] R. K. Das and R. V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, **2013**.

[12] A. S. Holehouse, R. K. Das, J. N. Ahad, et al. Cider: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical Journal*, 112 (1):16 – 21, **2017**.

[13] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, **2021**. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

[14] B. Strodel. Energy landscapes of protein aggregation and conformation switching in intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167182, **July 2021**. doi: 10.1016/j.jmb.2021.167182.

[15] K. M. Ruff and R. V. Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, **2021**. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2021.167208. URL https://www.sciencedirect.com/science/article/pii/S0022283621004411. From Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology.

[16] D. V. Laurents. Alphafold 2 and nmr spectroscopy: Partners to understand protein structure, dynamics and function. *Frontiers in Molecular Biosciences*, 9, **2022**. ISSN 2296-889X. doi: 10.3389/fmolb.2022.906437. URL https://www.frontiersin.org/articles/10.3389/fmolb.2022.906437.

[17] T. Ishida and K. Kinoshita. Prdos: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research*, 35:W460–W464, **07 2007**. ISSN 0305-1048. doi: 10.1093/nar/gkm363. URL https://doi.org/10.1093/nar/gkm363.

[18] B. Xue, R. L. Dunbrack, R. W. Williams, et al. Pondr-fit: A meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(4):996–1010, **2010**. ISSN 1570-9639. doi: https://doi.org/10.1016/j.bbapap.2010.01.011. URL `https://www.sciencedirect.com/science/article/pii/S1570963910000130`.

[19] G. Erdos, M. Pajkos, and Z. Dosztanyi. Iupred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, 49(W1):W297–W303, **05 2021**. ISSN 0305-1048. doi: 10.1093/nar/gkab408. URL `https://doi.org/10.1093/nar/gkab408`.

[20] B. Zhao and B. Xue. Decision-tree based meta-strategy improved accuracy of disorder prediction and identified novel disordered residues inside binding motifs. *International Journal of Molecular Sciences*, 19, **2018**.

[21] S. P. Humphrey and R. T. Williamson. A review of saliva: Normal composition, flow, and function. *The Journal of Prosthetic Dentistry*, 85(2):162–169, **2001**. ISSN 0022-3913. doi: https://doi.org/10.1067/mpr.2001.113778. URL `https://www.sciencedirect.com/science/article/pii/S0022391301540329`.

[22] K. Kavanagh and S. Dowd. Histatins: antimicrobial peptides with therapeutic potential. *Journal of Pharmacy and Pharmacology*, 56(3):285–289, **02 2010**. ISSN 0022-3573. doi: 10.1211/0022357022971. URL `https://doi.org/10.1211/0022357022971`.

[23] 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: Struct., Funct., Bioinf.*, 84(6):777–791, **2016**.

[24] D. Brewer, H. Hunter, and G. Lajoie. Nmr studies of the antimicrobial salivary peptides histatin 3 and histatin 5 in aqueous and nonaqueous solutions. *Biochemistry and Cell Biology*, 76(2-3):247–256, **1998**. doi: 10.1139/o98-066.

[25] P. A. Raj, E. Marcus, and D. K. Sukumaran. Structure of human salivary histatin 5 in aqueous and nonaqueous solutions. *Biopolymers*, 45(1):51–67, **1998**. doi: 10.1002/(SICI)1097-0282(199801)45:1<51::AID-BIP5>3.0.CO;2-Y.

[26] S. Melino, S. Rufini, M. Sette, et al. Zn2+ ions selectively induce antimicrobial salivary peptide histatin-5 to fuse negatively charged vesicles. identification and characterization of a zinc-binding motif present in the functional domain. *Biochemistry*, 38(30):9626–9633, **1999**.

[27] S. Puri and M. Edgerton. How does it kill?: Understanding the candidacidal mechanism of salivary histatin 5. *Eukaryotic Cell*, 13(8):958–964, **2014**.

[28] A. L. A. Ruissen, J. Groenink, E. J. Helmerhorst, et al. Effects of histatin 5 and derived peptides on candida albicans. *Biochem. J.*, 356(2):361–368, **2001**.

[29] K. Wróblewski, R. Muhandiram, A. Chakrabartty, and A. Bennick. The molecular interaction of human salivary histatins with polyphenolic compounds. *Eur. J. Biochem.*, 268(16):4384–4397, **2001**.

[30] A. Bennick. Interaction of plant polyphenols with salivary proteins. *Crit. Rev. Oral Biol. Med.*, 13:184–196, **2002**.

[31] L. Schnaider, A. Rosenberg, T. Kreiser, et al. Peptide self-assembly is linked to antibacterial, but not antifungal, activity of histatin 5 derivatives. *mSphere*, 5(2), **2020**. doi: 10.1128/mSphere.00021-20.

[32] S. B. Zimmerman and S. O. Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli. *J. Mol. Biol.*, 222 (3):599–620, **1991**.

[33] R. Balu, J. P. Mata, R. Knott, et al. Effects of crowding and environment on the evolution of conformational ensembles of the multi-stimuli-responsive intrinsically disordered protein, rec1-resilin: A small-angle scattering investigation. *Physical Chemistry B*, 120(27):6490–6503, **2016**.

[34] C. Szasz, A. Alexa, K. Toth, et al. Protein disorder prevails under crowded conditions. *Biochemistry*, 50(26):5834–5844, **2011**. doi: 10.1021/bi200365j.

[35] C. M. Miller, Y. C. Kim, and J. Mittal. Protein composition determines the effect of crowding on the properties of disordered proteins. *Biophys. J.*, 111(1):28–37, **2016**.

[36] A. V. Fonin, A. L. Darling, I. M. Kuznetsova, et al. Intrinsically disordered proteins in crowded milieu: when chaos prevails within the cellular gumbo. *Cellular and Molecular Life Sciences*, 75(21):3907–3929, **2018**.

[37] A. Bremer, M. Wolff, A. Thalhammer, and D. K. Hincha. Folding of intrinsically disordered plant lea proteins is driven by glycerol-induced crowding and the presence of membranes. *FEBS J.*, 284(6):919–936, **2017**. doi: https://doi.org/10.1111/febs.14023.

[38] E. A. Cino, M. Karttunen, and W.-Y. Choy. Effects of molecular crowding on the dynamics of intrinsically disordered proteins. *PLOS ONE*, 7(11):1–12, **11 2012**. doi: 10.1371/journal.pone.0049876. URL `https://doi.org/10.1371/journal.pone.0049876`.

[39] A. Bonucci, M. Palomino-Schättzlein, P. M. de Molina, et al. Crowding effects in the structure and dynamics of the intrinsically disordered nuckear chromatin protein

nupr1. *Frontiers in Molecular Biosciences*, 8, **July 2021**. doi: 10.3389/fmolb.2021.684622.

[40] S. L. Flaugh and K. J. Lumb. Effects of macromolecular crowding on the intrinsically disordered proteins c-fos and p27-kip1. *Biomacromolecules*, 2(2):538–540, **2001**. doi: 10.1021/bm015502z.

[41] D. P. Goldenberg and B. Argyle. Minimal effects of macromolecular crowding on an intrinsically disordered protein: A small-angle neutron scattering study. *Biophysical Journal*, 106:905–914, **February 2014**.

[42] A. Banks, S. Qin, K. L. Weiss, et al. Intrinsically disordered protein exhibits both compaction and expansion under macromolecular crowding. *Biophysical Journal*, 114:1067–1079, **March 2018**.

[43] G. L. Dignon, R. B. Best, and J. Mittal. Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annual Review of Physical Chemistry*, 71:53–75, **April 2020**. doi: 10.1146/annurev-physchem-071819-113553.

[44] E. W. Martin and A. S. Holehouse. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerging Topics in Life Sciences*, 4(3):307–329, **October 2020**. doi: 10.1042/ETLS20190164.

[45] T. J. Nott, E. Petsalaki, P. Farber, et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Molecular Cell*, 57(5): 936–947, **2015**.

[46] D. M. Mitrea, J. A. Cika, C. B. Stanley, et al. Self-interaction of npm1 modulates multiple mechanisms of liquid-liquid phase separation. *Nature Communications*, 9 (842), **2018**.

[47] Y. Lin, S. L. Currie, and M. K. Rosen. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *Journal of Biological Chemistry*, 292(46):19110–19120, **2017**.

[48] B. S. Schuster, G. L. Dignon, W. S. Tang, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences*, 117(21):11421–11431, **2020**. doi: 10.1073/pnas.2000223117. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2000223117`.

[49] R. M. Vernon and J. D. Forman-Kay. First-generation predictors of biological protein phase separation. *Current Opinion in Structural Biology*, 58:88–96, **2019**.

[50] A. K. Lancaster, A. Nutter-Upham, S. Lindquist, and O. D. King. Plaac: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*, 30(17):2501–2502, **05 2014**.

[51] Y. Wang, L. A. Benton, V. Singh, and G. J. Pielak. Disordered protein diffusion under crowded conditions. *J. Phys. Chem. Lett.*, 3(18):2703–2706, **2012**. doi: 10.1021/jz3010915.

[52] I. König, B. Schuler, A. Soranno, and D. Nettels. Impact of in-cell and in-vitro crowding on the conformations and dynamics of an intrinsically disordered protein. *Angewandte Chemie International Edition*, **2021**. doi: https://doi.org/10.1002/anie.202016804.

[53] S. J. Shire, Z. Shahrokh, and J. Liu. *Challenges in the Development of High Protein Concentration Formulations*, pages 131–147. Springer New York, New York, NY, **2010**. ISBN 978-0-387-76643-0. doi: 10.1007/978-0-387-76643-0_9. URL `https://doi.org/10.1007/978-0-387-76643-0_9`.

[54] W. Jiskoot, A. Hawe, T. Menzen, et al. Ongoing challenges to develop high concentration monoclonal antibody-based formulations for subcutaneous administration: Quo vadis? *Journal of Pharmaceutical Sciences*, 111(4):861–867, **2022**. ISSN 0022-3549. doi: https://doi.org/10.1016/j.xphs.2021.11.008. URL `https://www.sciencedirect.com/science/article/pii/S0022354921006146`.

[55] M. Pindrus, S. J. Shire, R. F. Kelley, et al. Solubility challenges in high concentration monoclonal antibody formulations: Relationship with amino acid sequence and intermolecular interactions. *Molecular Pharmaceutics*, 12(11):3896–3907, **2015**. doi: 10.1021/acs.molpharmaceut.5b00336. URL `https://doi.org/10.1021/acs.molpharmaceut.5b00336`. PMID: 26407030.

[56] T. L. Hill. *An Introduction to Statistical Thermodynamics*. Dover Publications, Inc, **1986**. ISBN 0-486-65242-4.

[57] S. Augé, P.-O. Schmit, C. A. Crutchfield, et al. Nmr measure of translational diffusion and fractal dimension. application to molecular mass measurement. *J. Phys. Chem. B*, 113(7):1914–1918, **2009**. doi: 10.1021/jp8094424.

[58] J. A. Marsh and J. D. Forman-Kay. Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical Journal*, 98(10):2383–2390, **2010**. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2010.02.006. URL `https://www.sciencedirect.com/science/article/pii/S0006349510002341`.

[59] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, et al. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, **1953**. doi: 10.1063/1.1699114.

[60] J. E. Amaro and J. N. Orce. Monte carlo simulation of covid-19 pandemic using planck's probability distribution. *Biosystems*, 218:104708, **2022**. ISSN 0303-2647. doi: https://doi.org/10.1016/j.biosystems.2022.104708. URL https://www.sciencedirect.com/science/article/pii/S0303264722000934.

[61] J. Bae, J. W. Park, and S. J. Lee. Limit surface/states searching algorithm with a deep neural network and monte carlo dropout for nuclear power plant safety assessment. *Applied Soft Computing*, 124:109007, **2022**. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2022.109007. URL https://www.sciencedirect.com/science/article/pii/S1568494622003271.

[62] I. Popova and J. K. Yau. Computing optimal portfolios of multi-assets with tail risk: the case of bitcoin. *Applied Economics Letters*, 0(0):1–9, **2022**. doi: 10.1080/13504851.2022.2074352. URL https://doi.org/10.1080/13504851.2022.2074352.

[63] S.-J. Hong and H. Najmi. Impact of high-speed rail on air travel demand between dallas and houston applying monte carlo simulation. *Journal of Air Transport Management*, 102:102222, **2022**. ISSN 0969-6997. doi: https://doi.org/10.1016/j.jairtraman.2022.102222. URL https://www.sciencedirect.com/science/article/pii/S0969699722000436.

[64] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, **2002**. ISBN 978-0-12-267351-1.

[65] M. Rao, C. Pangali, and B. Berne. On the force bias monte carlo simulation of water: methodology, optimization and comparison with molecular dynamics. *Molecular Physics*, 37(6):1773–1798, **1979**. doi: 10.1080/00268977900101321. URL https://doi.org/10.1080/00268977900101321.

[66] D. Bouzida, S. Kumar, and R. H. Swendsen. Efficient monte carlo methods for the computer simulation of biological molecules. *Phys. Rev. A*, 45:8894–8901, **Jun 1992**. doi: 10.1103/PhysRevA.45.8894. URL https://link.aps.org/doi/10.1103/PhysRevA.45.8894.

[67] P. Hebbeker, P. Linse, and S. Schneider. Optimal displacement parameters in monte carlo simulations. *Journal of Chemical Theory and Computation*, 12(4):1459–1465, **2016**. doi: 10.1021/acs.jctc.5b00797. URL https://doi.org/10.1021/acs.jctc.5b00797. PMID: 26950768.

[68] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, **1957**. doi: 10.1063/1.1743957. URL https://doi.org/10.1063/1.1743957.

[69] H. Berendsen and W. Van Gunsteren. Practical algorithms for dynamic simulations. *Molecular-dynamics simulation of statistical-mechanical systems*, pages 43–65, **1986**.

[70] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg. Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, **2015**. doi: 10.1021/ct5010406. URL `https://doi.org/10.1021/ct5010406`. PMID: 26574392.

[71] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18 (12):1463–1472, **1997**. doi: https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[72] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, **2007**.

[73] P. Atkins and R. Friedman. *Molecular Quantum Mechanics*. Oxford University Press, **2011**. ISBN 978-0-19-954142-3.

[74] W. D. Cornell, P. Cieplak, C. I. Bayly, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, **1995**. doi: 10.1021/ja00124a002.

[75] H. A. Lorentz. Ueber die anwendung des satzes vom virial in der kinetischen theorie der gase. *Annalen der Physik*, 248(1):127–136, **1881**. doi: https://doi.org/10.1002/andp.18812480110. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18812480110`.

[76] D. Berthelot. Sur le mélange des gaz. *Compt. Rendus*, 126(3), **1898**.

[77] W. F. van Gunsteren and H. J. Berendsen. Groningen molecular simulation (gromos) library manual. *Biomos, Groningen*, 24(682704):13, **1987**.

[78] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, et al. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, **1983**. doi: https://doi.org/10.1002/jcc.540040211. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540040211`.

[79] W. L. Jorgensen and J. Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, **1988**. doi: 10.1021/ja00214a001. URL `https://doi.org/10.1021/ja00214a001`. PMID: 27557051.

[80] J. Henriques, C. Cragnell, and M. Skepö. Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation*, 11(7):3420–3431, **2015**.

[81] S. Rauscher, V. Gapsys, M. J. Gajda, et al. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *Journal of Chemical Theory and Computation*, 11(11):5513–5524, **2015**. doi: 10.1021/acs.jctc.5b00736. URL `https://doi.org/10.1021/acs.jctc.5b00736`. PMID: 26574339.

[82] R. B. Best, W. Zheng, and J. Mittal. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124, **2014**. doi: 10.1021/ct500569b. URL `https://doi.org/10.1021/ct500569b`. PMID: 25400522.

[83] J. Mu, H. Liu, J. Zhang, et al. Recent force field strategies for intrinsically disordered proteins. *Journal of Chemical Information and Modeling*, 61(3):1037–1047, **2021**. doi: 10.1021/acs.jcim.0c01175. URL `https://doi.org/10.1021/acs.jcim.0c01175`. PMID: 33591749.

[84] P. S. Nerenberg and T. Head-Gordon. New developments in force fields for biomolecular simulations. *Current Opinion in Structural Biology*, 49:129–138, **April 2018**. doi: https://doi.org/10.1016/j.sbi.2018.02.002.

[85] R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, **2009**. doi: 10.1021/jp901540t. URL `https://doi.org/10.1021/jp901540t`. PMID: 19514729.

[86] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen. Improved peptide and protein torsional energetics with the opls-aa force field. *Journal of Chemical Theory and Computation*, 11(7):3499–3509, **2015**. doi: 10.1021/acs.jctc.5b00356. URL `https://doi.org/10.1021/acs.jctc.5b00356`. PMID: 26190950.

[87] E. Harder, W. Damm, J. Maple, et al. Opls3: A force field providing broad coverage of drug-like small molecules and proteins. *Journal of Chemical Theory and Computation*, 12(1):281–296, **2016**. doi: 10.1021/acs.jctc.5b00864. URL `https://doi.org/10.1021/acs.jctc.5b00864`. PMID: 26584231.

[88] S. Piana, K. Lindorff-Larsen, and D. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100(9):L47–L49, **2011**. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2011.03.051. URL `https://www.sciencedirect.com/science/article/pii/S0006349511004097`.

[89] A. D. MacKerell, M. Feig, and C. L. Brooks. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society*, 126(3): 698–699, **2004**. doi: 10.1021/ja036959e. URL https://doi.org/10.1021/ja036959e. PMID: 14733527.

[90] J. Huang, S. Rauscher, G. Nawrocki, et al. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73, **2017**.

[91] H. Liu, D. Song, H. Lu, et al. Intrinsically disordered protein-specific force field charmm36idpsff. *Chemical Biology & Drug Design*, 92(4):1722–1735, **2018**. doi: https://doi.org/10.1111/cbdd.13342. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cbdd.13342.

[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.*, 115(21): E4758–E4766, **2018**. doi: 10.1073/pnas.1800690115.

[93] E. A. Ponomarenko, E. V. Poverennaya, E. V. Ilgisonis, et al. The size of the human proteome: the width and depth. *International journal of analytical chemistry*, 2016, **2016**.

[94] R. Mecke. Das rotationsschwingungsspektrum des wasserdampfes. i. *Zeitschrift für Physik*, 81(5):313 – 331, **1933**. doi: 10.1007/BF01344550.

[95] J. Morgan and B. E. Warren. X-ray analysis of the structure of water. *The Journal of Chemical Physics*, 6(11):666–673, **1938**. doi: 10.1063/1.1750148. URL https://doi.org/10.1063/1.1750148.

[96] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids*, 101(1):219–260, **2002**. ISSN 0167-7322. doi: https://doi.org/10.1016/S0167-7322(02)00094-6. URL https://www.sciencedirect.com/science/article/pii/S0167732202000946. Molecular Liquids. Water at the New Millenium.

[97] A. V. Onufriev and D. A. Case. Generalized born implicit solvent models for biomolecules. *Annual Review of Biophysics*, 48(1):275–296, **2019**. doi: 10.1146/annurev-biophys-052118-115325. URL https://doi.org/10.1146/annurev-biophys-052118-115325. PMID: 30857399.

[98] W. L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *Journal of the American Chemical Society*, 103(2):335–340, **1981**. doi: 10.1021/ja00392a016. URL https://doi.org/10.1021/ja00392a016.

[99] A. V. Onufriev and S. Izadi. Water models for biomolecular simulations. *WIREs Computational Molecular Science*, 8(2):e1347, **2018**. doi: https://doi.org/10.1002/wcms.1347. URL `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1347`.

[100] 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*, 119(16):5113–5123, **2015**. doi: 10.1021/jp508971m. URL `https://doi.org/10.1021/jp508971m`. PMID: 25764013.

[101] 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. *Journal of Chemical Theory and Computation*, 12 (7):3407–3415, **2016**.

[102] A. Kurut, C. Dicko, and M. Lund. Dimerization of terminal domains in spiders silk proteins is controlled by electrostatic anisotropy and modulated by hydrophobic patches. *ACS Biomaterials Science & Engineering*, 1(6):363–371, **2015**. doi: 10.1021/ab500039q. URL `https://doi.org/10.1021/ab500039q`. PMID: 33445241.

[103] Y. Nozaki and C. Tanford. Enzyme structure. *Methods in Enzymology*, 11:715–734, **1967**.

[104] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An n log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, **1993**. doi: 10.1063/1.464397.

[105] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287, **1921**. doi: 10.1002/andp.19213690304. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19213690304`.

[106] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98–103, **Jul 1967**. doi: 10.1103/PhysRev.159.98. URL `https://link.aps.org/doi/10.1103/PhysRev.159.98`.

[107] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. McGraw-Hill, **1981**. ISBN 00702910 8X 9780070291089.

[108] A. Grossfield and D. M. Zuckerman. Chapter 2 quantifying uncertainty and sampling quality in biomolecular simulations. volume 5 of *Annual Reports in Computational Chemistry*, pages 23–48. Elsevier, **2009**. doi: https://doi.org/10.1016/S1574-1400(09)00502-7. URL `https://www.sciencedirect.com/science/article/pii/S1574140009005027`.

[109] A. Grossfield, P. N. Patrone, D. R. Roe, et al. Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]. *Living Journal of Computational Molecular Science*, 1(1):5067, **Oct. 2018**. doi: 10.33011/livecoms.1.1.5067. URL `https://livecomsjournal.org/index.php/livecoms/article/view/v1i1e5067`.

[110] S. R. R. Campos and A. M. Baptista. Conformational analysis in a multidimensional energy landscape: Study of an arginylglutamate repeat. *J. Phys. Chem. B*, 113(49): 15989 – 16001, **2009**. doi: 10.1021/jp902991u.

[111] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, **1963**. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(63)80023-6. URL `https://www.sciencedirect.com/science/article/pii/S0022283663800236`.

[112] S. A. Hollingsworth and P. A. Karplus. A fresh look at the ramachandran plot and the occurrence of standard structures in proteins. *BioMol Concepts*, 1(3-4):271–283, **2010**. doi: doi:10.1515/bmc.2010.022. URL `https://doi.org/10.1515/bmc.2010.022`.

[113] W. L. Jorgensen and J. Tirado-Rives. Monte carlo vs molecular dynamics for conformational sampling. *The Journal of Physical Chemistry*, 100(34):14508–14513, **1996**. doi: 10.1021/jp960880x.

[114] H. Yamashita, S. Endo, H. Wako, and A. Kidera. Sampling efficiency of molecular dynamics and monte carlo method in protein simulation. *Chemical Physics Letters*, 342(3):382 – 386, **2001**. doi: https://doi.org/10.1016/S0009-2614(01)00613-3.

[115] S. H. Northrup and J. A. McCammon. Simulation methods for protein structure fluctuations. *Biopolymers*, 19(5):1001–1016, **1980**. doi: https://doi.org/10.1002/bip.1980.360190506. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.1980.360190506`.

[116] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253 (5494):694–698, **1975**. doi: 10.1038/253694a0.

[117] P. Depa, C. Chen, and J. K. Maranas. Why are coarse-grained force fields too fast? a look at dynamics of four coarse-grained polymers. *The Journal of Chemical Physics*, 134(1):014903, **2011**. doi: 10.1063/1.3513365.

[118] A. Pak and G. Voth. Advances in coarse-grained modeling of macromolecular complexes. *Current Opinion in Structural Biology*, **2018**.

[119] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, et al. The power of coarse graining in biomolecular simulations. *WIREs Computational Molecular Science*, 4(3):225–248, **2014**. doi: https://doi.org/10.1002/wcms.1169.

[120] D. H. de Jong, G. Singh, W. F. D. Bennett, et al. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.*, 9(1):687–697, **2013**.

[121] T. Bereau and M. Deserno. Generic coarse-grained model for protein folding and aggregation. *The Journal of Chemical Physics*, 130(23):235106, **2009**. doi: 10.1063/1.3152842.

[122] M. van der Klok, M. Dekker, E. Van der Giessen, and P. R. Onck. A 4bpa coarse-grained molecular dynamics study on the aggregation of polyglutamine. *Biophysical Journal*, 120(3, Supplement 1):28a, **2021**. ISSN 0006-3495. doi: https://doi.org/10.1016/j.bpj.2020.11.428. URL `https://www.sciencedirect.com/science/article/pii/S0006349520313357`.

[123] G. L. Dignon, W. Zheng, Y. C. Kim, et al. Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Computational Biology*, 14(1):1–23, **January 2018**.

[124] U. Baul, D. Chakraborty, M. L. Mugnai, et al. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. *J. Phys. Chem. B*, 123(16):3462–3474, **2019**.

[125] M. Neri, C. Anselmi, M. Cascella, et al. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.*, 95:218102, **Nov 2005**. doi: 10.1103/PhysRevLett.95.218102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.95.218102`.

[126] P. Kar and M. Feig. Chapter five - recent advances in transferable coarse-grained modeling of proteins. In T. Karabencheva-Christova, editor, *Biomolecular Modelling and Simulations*, volume 96 of *Advances in Protein Chemistry and Structural Biology*, pages 143–180. Academic Press, **2014**. doi: https://doi.org/10.1016/bs.apcsb.2014.06.005. URL `https://www.sciencedirect.com/science/article/pii/S1876162314000066`.

[127] D. Schneidman-Duhovny, M. Hammel, and A. Sali. Foxs: a web server for rapid computation and fitting of saxs profiles. *Nucleic Acids Research*, 38:W540–W544, **05 2010**. ISSN 0305-1048. doi: 10.1093/nar/gkq461. URL `https://doi.org/10.1093/nar/gkq461`.

[128] D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, and A. Sali. Accurate saxs profile computation and its assessment by contrast variation experiments. *Biophysical Journal*, 105:962–974, **2013**. doi: 10.1016/j.bpj.2013.07.020.

[129] S. Grudinin, M. Garkavenko, and A. Kazennov. *Pepsi-SAXS*: an adaptive method for rapid and accurate computation of small-angle x-ray scattering profiles. *Acta Crystallographica Section D*, 73(5):449–464, **May 2017**. doi: 10.1107/S2059798317005745. URL `https://doi.org/10.1107/S2059798317005745`.

[130] J. Henriques, L. Arleth, K. Lindorff-Larsen, and M. Skepö. On the calculation of saxs profiles of folded and intrinsically disordered proteins from computer simulations. *Journal of Molecular Biology*, 430(16):2521–2539, **2018**. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2018.03.002. URL `https://www.sciencedirect.com/science/article/pii/S0022283618301232`. Intrinsically Disordered Proteins.

[131] E. J. Maginn, R. A. Messerly, D. J. Carlson, et al. Best practices for computing transport properties 1. self-diffusivity and viscosity from equilibrium molecular dynamics [article v1.0]. *Living Journal of Computational Molecular Science*, 1(1):6324, **Dec. 2018**. doi: 10.33011/livecoms.1.1.6324. URL `https://livecomsjournal.org/index.php/livecoms/article/view/v1i1e6324`.

[132] I.-C. Yeh and G. Hummer. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B*, 108(40):15873–15879, **2004**. doi: 10.1021/jp0477147.

[133] A. Ortega, D. Amoros, and J. G. de la Torre. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.*, 101(4):892–898, **August 2011**. doi: 10.1016/j.bpj.2011.06.046.

[134] V. Bloomfield, W. O. Dalton, and K. E. Van Holde. Frictional coefficients of multisubunit structures. i. theory. *Biopolymers*, 5(2):135–148, **1967**. doi: https://doi.org/10.1002/bip.1967.360050202. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.1967.360050202`.

[135] G. Nagy, M. Igaev, N. C. Jones, et al. Sesca: Predicting circular dichroism spectra from protein molecular structures. *Journal of Chemical Theory and Computation*, 15 (9):5087–5102, **2019**. doi: 10.1021/acs.jctc.9b00203. URL `https://doi.org/10.1021/acs.jctc.9b00203`. PMID: 31402660.

[136] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, **1983**. doi: 10.1002/bip.360221211. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360221211`.

[137] G. Nagy and C. Oostenbrink. Dihedral-based segment identification and classification of biopolymers i: Proteins. *Journal of Chemical Information and Modeling*, 54 (1):266–277, **2014**. doi: 10.1021/ci400541d. URL https://doi.org/10.1021/ci400541d. PMID: 24364820.

[138] A. Guinier. La diffraction des rayons x aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys.*, 11(12):161–237, **1939**. doi: 10.1051/anphys/193911120161. URL https://doi.org/10.1051/anphys/193911120161.

[139] V. Receveur-Bréchot and D. Durand. How random are intrinsically disordered proteins? a small angle scattering perspective. *Current Protein and Peptide Science*, 13: 55–75, **2012**.

[140] W. Zheng and R. B. Best. An extended guinier analysis for intrinsically disordered proteins. *Journal of Molecular Biology*, 430:2540–2553, **2018**.

[141] D. Durand, C. Vivès, D. Cannella, et al. Nadph oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers. *Journal of Structural Biology*, 169(1):45 – 53, **2010**. ISSN 1047-8477. doi: https://doi.org/10.1016/j.jsb.2009.08.009. URL http://www.sciencedirect.com/science/article/pii/S1047847709002391.

[142] D. Johansen, J. Trewhella, and D. P. Goldenberg. Fractal dimension of an intrinsically disordered protein: Small-angle x-ray scattering and computational study of the bacteriophage λ n protein. *Protein Science*, 20(12):1955–1970, **2011**. doi: 10.1002/pro.739. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.739.

[143] M. Bee. *Quasielastic neutron scattering*. Adam Hilger, **1988**. ISBN 0-85274-371-8.

[144] K. Singwi and A. Sjölander. Diffusive motions in water and cold neutron scattering. *Phys. Rev.*, 119(3):863–871, **Aug 1960**. doi: 10.1103/PhysRev.119.863.

[145] Chapter 6 - circular dichroism spectroscopy for protein characterization: Biopharmaceutical applications. In D. J. Houde and S. A. Berkowitz, editors, *Biophysical Characterization of Proteins in Developing Biopharmaceuticals*, pages 109 – 137. Elsevier, Amsterdam, **2015**. ISBN 978-0-444-59573-7. doi: https://doi.org/10.1016/B978-0-444-59573-7.00006-3.

[146] S. M. Kelly, T. J. Jess, and N. C. Price. How to study proteins by circular dichroism. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1751(2):119 – 139, **2005**. ISSN 1570-9639. doi: https://doi.org/10.1016/j.bbapap.2005.06.005. URL http://www.sciencedirect.com/science/article/pii/S1570963905001792.

[147] S. G. Ramalli, A. J. Miles, R. W. Janes, and B. Wallace. The pcddb (protein circular dichroism data bank): A bioinformatics resource for protein characterisations and methods development. *Journal of Molecular Biology*, 434(11): 167441, **2022**. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2022.167441. URL `https://www.sciencedirect.com/science/article/pii/S0022283622000018`. Computation Resources for Molecular Biology.

[148] J. G. Lees, A. J. Miles, F. Wien, and B. A. Wallace. A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics*, 22(16):1955–1962, **06 2006**. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl327. URL `https://doi.org/10.1093/bioinformatics/btl327`.

[149] 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*, 27(12):1630–1636, **04 2011**. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr234. URL `https://doi.org/10.1093/bioinformatics/btr234`.

[150] J. Tolchard, S. J. Walpole, A. J. Miles, et al. The intrinsically disordered tarp protein from chlamydia binds actin with a partially preformed helix. *Scientific Reports*, 8(1): 1960, **2018**. doi: 10.1038/s41598-018-20290-8.

[151] A. Micsonai, F. Wien, E. Bulyaki, et al. Bestsel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Research*, 46(W1):W315–W322, **2018**. doi: 10.1093/nar/gky497.

[152] J. Huang and A. D. MacKerell. Force field development and simulations of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 48: 40 – 48, **2018**. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2017.10.008. URL `http://www.sciencedirect.com/science/article/pii/S0959440X17301148`. Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective.

[153] R. V. Pappu, X. Wang, A. Vitalis, and S. L. Crick. A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Arch. Biochem. Biophys.*, 469 (1):132–141, **2008**. doi: 10.1016/j.abb.2007.08.033.

[154] C. Cragnell, E. Rieloff, and 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):2478–2492, **2018**.

[155] D. Roberts, R. Keeling, M. Tracka, et al. Specific ion and buffer effects on protein–protein interactions of a monoclonal antibody. *Mol. Pharmaceutics*, 12(1):179–193,

2015. doi: 10.1021/mp500533c. URL https://doi.org/10.1021/mp500533c. PMID: 25389571.

[156] F. Xie, M. Turesson, M. Jansson, et al. A simple and versatile implicit solvent model for polyethylene glycol in aqueous solution at room temperature. *Polymer*, 84:132–137, **2016**. ISSN 0032-3861. doi: https://doi.org/10.1016/j.polymer.2015.12.034. URL https://www.sciencedirect.com/science/article/pii/S003238611530447X.

[157] M. Javanainen, H. Martinez-Seara, and I. Vattulainen. Excessive aggregation of membrane proteins in the martini model. *PLoS One*, 12(11):1–20, **11 2017**. doi: 10.1371/journal.pone.0187936.

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

[159] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun. Advanced ensemble modelling of flexible macromolecules using x-ray solution scattering. *IUCrJ*, 2(2):207–217, **March 2015**. doi: 10.1107/S205225251500202X.

[160] A. Irbäck and S. Mohanty. Profasi: A monte carlo simulation package for protein folding and aggregation. *J. Comput. Chem.*, 27(13):1548–1555, **2006**.

[161] A. Sagar, C. M. Jeffries, M. V. Petoukhov, et al. Comment on the optimal parameters to derive intrinsically disordered protein conformational ensembles from small-angle x-ray scattering data using the ensemble optimization method. *Journal of Chemical Theory and Computation*, 17(4):2014–2021, **2021**. doi: 10.1021/acs.jctc.1c00014. URL https://doi.org/10.1021/acs.jctc.1c00014. PMID: 33725442.

[162] K. Lindorff-Larsen, S. Piana, K. Palmo, et al. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Struct., Funct., Bioinf.*, 78(8):1950–1958, **2010**. doi: 10.1002/prot.22711. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22711.

[163] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys.*, 123(23):234505, **2005**. doi: 10.1063/1.2121687. URL https://doi.org/10.1063/1.2121687.

[164] 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.*, 17(10):6634–6646, **2021**. doi: 10.1021/acs.jctc.1c00408.

[165] U. Shrestha, J. Smith, and L. Petridis. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun. Biol.*, 4 (1):243, **2021**. doi: 10.1038/s42003-021-01759-1.

[166] S. Fujiwara, K. Araki, T. Matsuo, et al. Dynamical Behavior of Human α-Synuclein studied by Quasielastic Neutron Scattering. *PloS One*, 11(4):e0151447, **2016**.

[167] K. Rogers. What is the difference between a peptide and a protein? `https://www.britannica.com/story/what-is-the-difference-between-a-peptide-and-a-protein`, . Accessed: 2022-07-14.

Scientific publications

Author contributions

Paper I: Evaluating models of varying complexity of crowded intrinsically disordered protein solutions against SAXS

I performed and analysed all of the simulations, participated in experimental work, participated in analysis of experiments, and co-wrote the article. Responsible for submission and revision process.

Paper II: The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions

I performed and analyzed all of the simulations, participated in experimental work, participated in analysis of experiments, and co-wrote the article. Responsible for submission and revision process.

Paper III: Self-Diffusive Properties of the Intrinsically Disordered Protein Histatin 5 and the Impact of Crowding Thereon: A Combined Neutron Spectroscopy and Molecular Dynamics Simulation Study

I performed and analysed all of the simulations, participated in experimental work, participated in analysis of experiments, and co-wrote the article. Responsible for submission and revision process.

Paper IV: The crowding effect using neutral crowders on Histatin 5. Computer simulations in combination with X-ray and dynamic light scattering

I performed and analysed all of the simulations, participated in DLS measurements, participated in analysis of all experiments, and co-wrote the manuscript.

Paper V: Comparative performance of coarse-grained IDP models at different resolutions

I participated in the design of the study, performed and analysed all simulations, co-wrote the manuscript.

Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University

ISBN 978-91-7422-898-4