LUND UNIVERSITY

Modeling and simulation of intrinsically disordered proteins

Henriques, Joao

2016

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

Link to publication

Citation for published version (APA):
Henriques, J. (2016). *Modeling and simulation of intrinsically disordered proteins*. [Doctoral Thesis (compilation), Computational Chemistry]. Lund University, Faculty of Science, Department of Chemistry, Theoretical Chemistry.

Total number of authors:
1

Creative Commons License:
CC BY-NC-ND

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

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

Modeling and simulation of intrinsically disordered proteins

THEORETICAL CHEMISTRY | FACULTY OF SCIENCE | LUND UNIVERSITY
JOÃO HENRIQUES

Modeling and simulation of intrinsically disordered proteins

João Henriques

Division of Theoretical Chemistry
Lund University, Sweden

LUND
UNIVERSITY

Doctoral Thesis

The faculty opponent is Robert B. Best, Investigator at the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, United States.

The thesis will be publicly defended on Friday, the 9th of December 2016, at 13:00, in lecture hall B of the Center for Chemistry and Chemical Engineering (Kemicentrum), Lund.

Organization
LUND UNIVERSITY

Center for Chemistry and Chemical Engineering
P. O. Box 124
SE–221 00 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2016-12-09

Sponsoring organization
The OMM Linnaeus center at Lund University (Swedish Research Council)

Author(s)
João Henriques

Title and subtitle
Modeling and simulation of intrinsically disordered proteins:

Abstract

This work is primarily about the development, validation and application of computer simulation models for intrinsically disordered proteins, both in solution and in the presence of uniformly charged, ideal surfaces. The models in question are either coarse-grained or atomistic in nature, and their applications are dependent on the specific purpose of each study. Both, Metropolis Monte Carlo and molecular dynamics simulations were employed to execute them.

In regard to the coarse-grained models, it was found that a simple physical model can be used to mimic the properties of flexible proteins, helping to understand how and why these proteins adsorb to surfaces under certain conditions. The same model later shown that two disordered proteins from different sources (saliva and milk) possess similar structural and thermodynamic properties in solution and when adsorbed to surfaces, thus being hypothesized that it may be possible to use one of them as a substitute for the other under a pharmaceutical context.

After a first indication that the atomistic models used until recently for the simulation of well-folded proteins may not be applicable to their disordered counterparts, it was then confirmed - by evaluating several such models against experimental evidence - that these models do indeed produce overly collapsed IDP conformational ensembles. New models, favoring protein–water over protein–protein interactions, were then shown to effectively produce more extended conformations, which are in much better agreement with each other and with experimental evidence. One of the new atomistic models was then used to perform the structural characterization of a disordered peptide conjugated to a small molecule, which has been shown to possess promising therapeutical applications. The value of computer simulations is well illustrated in this study, as the insight obtainable from experiment was limited and it was only through the analysis of the simulations that a possible link between the average conjugate structure and its increased antifungal activity is established.

DOKUMENTDATABLAD enl SIS 61 41 21

Key words

Intrinsically disordered proteins, coarse-grained models, atomistic models, Metropolis Monte Carlo simulations, molecular dynamics simulations, sampling, conformational analysis, charge regulation, surface adsorption

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN

978-91-7422-489-4 (print)
978-91-7422-490-0 (pdf)

Recipient's notes

Number of pages
170

Price

Security classification

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

Signature

Date November 1, 2016

Modeling and simulation of intrinsically disordered proteins

João Henriques

Division of Theoretical Chemistry
Lund University, Sweden

LUND
UNIVERSITY

Front cover illustration: "Grunge Flat Cut Stone Texture", taken from `wildtextures.com`

Funding information: The OMM Linnaeus center at Lund University (Swedish Research Council)

© João Henriques 2016

Faculty of Science, Center for Chemistry and Chemical Engineering

ISBN: 978-91-7422-489-4 (print)
ISBN: 978-91-7422-490-0 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2016

All work and no play makes Jack a dull boy
All work and no play makes Jack a dull boy
All work and no play mmakes Jack a dull boy
v All work and no Play ma es Jack a dull boy
Allworkand noplaymakesJack a dull boy
All work and no play makes Jack a dullboy.
All work and no play makes Jack a dull boy
All work and notplay makes Jack a dull boy

Preface

A doctoral thesis at Lund University either takes 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 authors. This thesis belongs to the latter category and it covers a large portion of the research I conducted during the past five years of my academic trajectory. All the papers included in this work are, in one way or another, devoted to the development and/or application of computer models for the simulation of intrinsically disordered proteins; hence the thesis title. The investigation of molecular phenomena by computer simulation is one of my passions, and my interest in protein disorder is partially derived from the challenge it presents to the otherwise established panorama of protein modeling and simulation. This work represents my modest contribution to understanding the limitations of current protein models and the development, validation and application of new ones.

This journey was long, hard and particularly eventful. The young and carefree partygoer that arrived in Lund on the 10th of January 2012, bears little resemblance to the 30 year old family man currently writing this thesis. Still, I would not have it any other way, and it was all a part of a much needed maturing process as a researcher and as an individual. All in all, I feel very privileged to have been able to undertake my doctoral studies at such a prestigious institution, surrounded by a group of immensely talented and knowledgeable people. In particular, I would like to express my gratitude to Marie, my main supervisor and mentor, for all your patience and support. It cannot have been easy to guide someone as headstrong (and at times obstinate) as myself. Yet, we have always maintained a very healthy mutual understanding and your experience as a researcher and as a parent have been absolutely invaluable to me. To Mikael, my co-supervisor, thank you for all the fruitful discussions and technical advice. Your mastery of simulation software programming and development is second to none. Since day one, it has been a great advantage to be in a research group where simulation and experiment go hand in hand. Therefore, I would like to thank my co-authors, whose experimental results were essential for the validation of my work. To Stephanie, who apart from being my colleague and co-author, was also my student/supervisee on three separate occasions, thank you for making me realize how much I enjoy teaching. To all my other colleagues and collaborators, I would like to say that it was an enriching experience to work with such a skilled and diverse group of researchers. To end - and I assume that this goes without saying - this thesis would not have materialized without the extensive support of my family and friends. In particular, my partner Maria, who has always been there for me, specially during rough times. I am greatly indebted to you!

João Henriques
Lund, October 2016

Contents

List of publications iii

Popular scientific summary in English v

Resumo simplificado em Português vii

Modeling and simulation of intrinsically disordered proteins: 1

1 An introduction to protein disorder 1

2 Theoretical background 5
- 2.1 Statistical thermodynamics 5
- 2.2 Classical statistical mechanics 8
- 2.3 Intermolecular interactions 9
  - 2.3.1 Charge–charge 10
  - 2.3.2 Charge–dipole 12
  - 2.3.3 Dipole–dipole 13
  - 2.3.4 Charge–non-polar 14
  - 2.3.5 Dipole–non-polar 15
  - 2.3.6 Non-polar–non-polar 15
  - 2.3.7 van der Waals forces 17
  - 2.3.8 Hydrogen bond 18

3 Modeling 21
- 3.1 Coarse-grained modeling of flexible proteins 21
- 3.2 Atomistic modeling of proteins 25
- 3.3 Explicit water models 27
- 3.4 Considerations about force fields 29

4 Monte Carlo simulations 31
- 4.1 The Metropolis method 31
- 4.2 Trial moves 32
- 4.3 Example of a basic algorithm 34

5 Molecular dynamics simulations 37
- 5.1 Equations of motion 37
- 5.2 Finite difference methods 38

6 Simulation techniques 41
- 6.1 Periodic boundary conditions 41
- 6.2 Potential truncation and the minimum image convention . . 43
- 6.3 Long-range force handling 44
- 6.4 Neighbor lists 45
- 6.5 Bond and angle constraints 46
- 6.6 Constant temperature and pressure 46

7 Simulation analyses 49
- 7.1 Size, shape and stiffness 49
- 7.2 Total charge and charge capacitance 50
- 7.3 Principal component analysis 51
- 7.4 Small-angle X-ray scattering 52

8 Summary of results and outlook 57
- 8.1 Paper I 57
- 8.2 Paper II 58
- 8.3 Paper III 58
- 8.4 Paper IV 59
- 8.5 Paper V 59
- 8.6 Paper VI 60
- 8.7 Outlook 61

9 References 63

Scientific publications **69**

Author contributions 69

Paper I: Role of histidine for charge regulation of unstructured peptides at interfaces and in bulk 71

Paper II: A coarse-grained model for flexible (phospho)proteins: adsorption and bulk properties 85

Paper III: *In silico* physicochemical characterization and comparison of two intrinsically disordered phosphoproteins: β-casein and acidic PRP-1 95

Paper IV: Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment 109

Paper V: Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models 123

Paper VI: Structural characterization of Histatin 5-spermidine conjugates: a combined experimental and theoretical study 135

List of publications

This thesis is based on the following publications:

I **Role of histidine for charge regulation of unstructured peptides at interfaces and in bulk**

A. Kurut, **J. Henriques**, J. Forsman, M. Skepö and M. Lund
Proteins: Structure, Function, and Bioinformatics, 2014, 82, pp. 657–667

II **A coarse-grained model for flexible (phospho)proteins: adsorption and bulk properties**

J. Henriques and M. Skepö
Food Hydrocolloids, 2015, 43, pp. 473–480

III ***In silico* physicochemical characterization and comparison of two intrinsically disordered phosphoproteins: β-casein and acidic PRP-1**

J. Henriques, S. Jephthah and M. Skepö
Food Hydrocolloids, 2016, 56, pp. 360–371

IV **Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment**

J. Henriques, C. Cragnell and M. Skepö
Journal of Chemical Theory and Computation, 2015, 11, pp. 3420–3431

V **Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models**

J. Henriques and M. Skepö
Journal of Chemical Theory and Computation, 2016, 12, pp 3407–3415

VI **Structural characterization of Histatin 5-spermidine conjugates: a combined experimental and theoretical study**

S. Jephthah, **J. Henriques**, C. Cragnell, S. Puri, M. Edgerton, and M. Skepö
Submitted to Macromolecules

All papers are reproduced with permission of their respective publishers.

Popular scientific summary in English

Proteins are large, complex molecules that play many critical roles in the body and are required for the structure, function, and regulation of the body's tissues and organs. In very simple terms, a protein can be defined as a linear chain of subunits called amino acid residues[1]. The individual amino acid residues are sequentially bonded together by peptide bonds. There are 20 standard amino acids in nature and their specific order of appearance in a protein chain is thought to determine its structure and function. For a very long time it was assumed that the structure of a protein and its function were mutually inclusive, and exceptions to the norm were often either "swept under the rug" or branded as mere curiosities of little relevance. Around the turn of the millennium, mounting evidence of structural disorder in a considerable amount of otherwise perfectly functional proteins lead to a change in paradigm. It is now known that structural disorder is not only abundant in all species, as it is also an advantage for proteins involved in functions which benefit from structural malleability.

Despite the considerable, recent interest in the study of intrinsically disordered proteins (likely due to their implication in a number of human diseases), the lack of a well-defined structure represents a substantial obstacle to their structural characterization by classic, high-resolution experimental methods. Some lower resolution methods can provide information about the average shape and size of the collection of structures that a disordered protein can attain in solution. However, computational methods are generally necessary to aid in interpreting and complementing the information that can be obtained from experimental data. One of such methods is the computer simulation of proteins, where a computer model of the protein[2] is run for a certain period of time, in order to observe and register the most relevant spatial arrangements which the protein may adopt and freely convert between. It is from this collection of arrangements that interesting structural and thermodynamic properties can be calculated, making computer simulations very powerful tools.

The papers[3] included in this thesis deal with the development, validation and application of computer simulation models for flexible and disordered proteins, both in solution and at interfaces[4]. In Paper II it was found that a simple physical model[5] can be used to mimic the properties of flexible proteins, helping to understand how and why these proteins adsorb to surfaces under certain conditions. In Paper III, the same simple model shown that two disordered proteins from different sources (saliva and milk) have very similar properties in solution and when adsorbed to surfaces. Thus, it was hypothesized that it may be possible to use one of them as a substitute for the other under a pharmaceutical context. Paper I was the catalyst for a series of studies (Papers IV – VI) involving more detailed protein models[6]. Among other things, this study provided an indication that the atomistic models used until then, for the simulation of proteins with well-defined structures, may not be applicable to their disordered counterparts. This was later confirmed in Paper IV, by evaluating several such models against experimental evidence. A similar evaluation was conducted for two new independent approaches developed with disordered proteins in mind. The results (presented in Papers IV and V) were shown to be in excellent agreement with each other and with experiment, which represents a considerable step forward in the search for accurate and predictive models for the simulation of disordered proteins. Finally, in Paper VI, one of the new atomistic models was used to perform the structural characterization of a disordered peptide conjugated to a small molecule, which has been shown to possess promising therapeutical applications. The value of computer simulations is well illustrated in this study, as the insight obtainable from experiment is limited[7] and it is only through the analysis of the simulations that a possible link between the average conjugate structure and its increased antifungal activity was established.

---

[1] In stricter terms, this linear chain of amino acid residues is called a polypeptide, and a protein contains *at least* one long polypeptide. Short polypeptides, containing less than 20–30 residues, are often referred to as peptides instead of proteins.

[2] A computer model is a combination of algorithms and equations used to capture the behavior of the system being modeled, i.e., a protein in this case.

[3] A scientific paper is a written (and preferentially published) report describing original research results.

[4] E.g., a charged surface.

[5] Where the protein is described at an intermediate level, i.e., the amino acid residues, which contain several atoms on their own, are simply represented as a single sphere. As such, the complete protein ends up resembling a pearl necklace, where each amino acid residue is a pearl bead. This type of model is usually referred to as a coarse-grained model.

---

[6] These models describe the protein at an atomistic level, and are thus usually referred to as "atomistic" models.

[7] Only the peptide part of the conjugate is detected by the instrument.

Resumo simplificado em Português

As proteínas são macromoléculas biológicas constituídas por uma ou mais cadeias de aminoácidos. Encontram-se presentes em todos os seres vivos e participam em praticamente todos os processos celulares, desempenhando um vasto conjunto de funções no organismo. Na natureza, existem 20 aminoácidos principais e a sua sequência, i.e., a ordem de aparição dos mesmos ao longo da cadeia, é tida como sendo determinante para a estrutura e função de uma proteína. Durante um longo período de tempo pensou-se que função de uma proteína estaria intimamente relacionada com a existência de uma única estrutura bem definida (a estrutura nativa), e eventuais excepções à regra foram consistentemente ignoradas ou tratadas como meras curiosidades sem relevância biológica. Por volta do virar do milénio, começou a ser evidente que um número considerável de proteínas possuem desordem estrutural parcial ou total. No entanto, estas proteínas mantêm-se perfeitamente funcionais no organismo, o que obrigou a uma mudança de paradigma. Hoje em dia é sabido que a desordem estrutural é abundante em todas as espécies, sendo altamente vantajosa para proteínas cujas funções requerem maleabilidade estrutural.

Apesar da forte aposta no estudo das proteínas intrinsecamente desordenadas (devido ao seu envolvimento em várias doenças graves), a ausência de uma estrutura estável e bem definida representa um grande obstáculo à sua caracterização através dos métodos experimentais de alta definição mais comuns. Alguns métodos de menor definição podem ser usados de modo a obter uma ideia do formato e da dimensão média de uma proteína desordenada em solução. No entanto, este tipo de informação é relativamente limitado, e é cada vez mais práctica comum recorrer a métodos computacionais para complementar e ajudar a interpretar estes dados experimentais. Na simulação de proteínas, um modelo computacional - i.e, um conjunto de algorítmos e equações desenvolvidos de modo a capturar o comportamento do sistema em estudo - é executado durante um período de tempo necessário para gerar (e guardar) o conjunto de conformações representativas do sistema em causa. É a partir deste conjunto de conformações que um número de importantes características estruturais e propriedades termodinâmicas podem ser calculadas, o que faz com que simulação computacional seja uma ferramenta muito poderosa e importante.

O trabalho incluído nesta tese de doutoramento foca-se no desenvolvimento, validação e aplicação de diferentes modelos para a simulação computacional de proteínas intrinsecamente desordenadas, tanto em solução como em contacto com superfícies. De forma geral, os resultados aqui apresentados representam uma contribuição positiva para o avanço desta área do conhecimento científico, demonstrando de forma clara a utilidade e precisão dos modelos mais recentes.

Modeling and simulation of intrinsically disordered proteins:

1 An introduction to protein disorder

Until around the turn of the millennium, evidence steadily accumulated that a well-defined structure is the prerequisite of protein function. This *structure–function paradigm* became so deeply ingrained that most biology and biochemistry textbooks relied, almost exclusively, in this notion in order to explain biological phenomena at the molecular level (Tompa 2012). However, after the first experimental observations of disorder in a few dozen proteins, it quickly became apparent that proteins and protein domains whose native and functional state is intrinsically unstructured[8] are common across the three domains of life[9], with special incidence in eukaryotic proteomes[10]. This realization forced a transition in paradigm (Wright & Dyson 1999), and it is now widely accepted that "unstructural" biology is an integral part of molecular biology (Tompa 2011).

A large collection of names can be found in the literature when describing disordered proteins[11]. In this work, the terms "intrinsically disordered" and "flexible" are preferred, with the latter being employed almost exclusively when referring to coarse-grained models. Notice that Dunker et al. (2013)[12] support the use of a single common term to describe these proteins, and "intrinsically disordered proteins" or IDPs is suggested as the most appropriate nomenclature.

---

[8] In part or in full.

[9] Bacteria, Archaea and Eukarya.

[10] More than one third of eukaryotic proteins have been shown to contain intrinsically disordered regions of over 30 amino acid residues in length (Ward et al. 2004).

[11] Floppy, pliable, rheomorphic, flexible, mobile, partially folded, natively denatured, natively unfolded, natively disordered, intrinsically unstructured, intrinsically denatured, intrinsically unfolded, intrinsically disordered, vulnerable, chameleon, malleable, 4D, protein clouds, dancing proteins, proteins waiting for partners, and several other names often representing different combinations of "natively/naturally/inherently/intrinsically" with "unfolded/unstructured/disordered/denatured" (Dunker et al. 2013).

[12] Whose authors are some of the most prominent figures in protein disorder research.

Table 1: List of IDPs studied in the publications included in this work and some of their fundamental physicochemical properties. Histatin 5_{4-15} corresponds to the active segment of Histatin 5, i.e., the contiguous amino acid sequence from residues 4 to 15. It does not exist *in vivo*. The percentage of disorder-promoting residues is calculated according to the key disorder-promoting amino acids residues as listed by Williams et al. (2001). pI is the isoelectric point, that is, the pH value for which the protein has a net charge of zero. FCR and NCPR stand for *fraction of charged residues* and *net charge per residue*, respectively, as defined by Das et al. (2015).

	Histatin 5	Histatin 5_{4-15}	β-casein	PRP-1
Paper(s)	I, IV–VI	VI	II, III	III
Organism	*Homo sapiens*	–	*Bos taurus*	*Homo sapiens*
Uniprot ID	P15516	P15516	P02666	P02810
Seq. position	20–43	23–34	16–224	17–166
Sequence length	24	12	209	150
% disorder-promoting res.	54.2	58.3	54.6	84.7
% titrable res.	66.7	66.7	21.1	17.3
% hydrophilic res.	83.3	75.0	47.5	66.0
% hydrophobic res.	16.7	25.0	52.6	34.0
pI	10.3	11.2	5.2	4.6
Net charge (pH 7)	5.4	5.1	−7.6	−7.0
FCR	0.40	0.44	0.18	0.16
NCPR	0.23	0.44	0.04	0.05

This class of proteins is characterized by broad structural, dynamic and functional characteristics. More specifically, IDPs can be classified into distinct conformational classes based on their amino acid compositions, and - from a functional point of view - IDPs are often implicated in important cellular processes that include cell division and signaling, intracellular transport, bacterial translocation, cell mechanics, protein degradation, posttranscriptional regulation, and cell cycle control (Das et al. 2015). These functions typically require binding to multiple partners, where high-specificity and/or low-affinity interactions play a crucial role, and they are only possible due to the intrinsic disorder of these proteins. Just as with their folded counterparts, numerous IDPs are also associated with human diseases, including cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, and diabetes (Uversky et al. 2008). The association with pathology is, no doubt, one of the main reasons behind the fast-growing interest in these proteins.

Curiously, the IDPs studied in the publications included in this work (Histatin 5, β-casein and PRP-1; see Table 1) are not believed to be the cause of any known disease(s). They may, however, present potential pharmaceutical applications. In particular, Histatin 5 (Oppenheim et al. 1988) and its active fragment Histatin 5_{4-15} (Wei & Bobek 2005), belong to the Histatin protein family, which is composed of closely related salivary and histidine-rich IDPs with a myriad of functions, from the maintenance of oral health to the integrity of the tooth surface. However, it is their high efficacy against fungal infections, namely the blastospore and the germinated form of *Candida albicans* (Xu et al. 1991), that is their most interesting characteristic, given the possibility of using enhanced Histatin 5 variants in therapeutic contexts. The proline-rich protein 1 or PRP-1 (Wong & Bennick 1980), is another salivary IDP with interesting pharmaceutical applications. It belongs to the proline-rich protein family (PRPs) (Hay et al. 1988), which are involved in the remineralization of the teeth and in tissue coating, being thus essential for the maintenance of the tooth enamel (Bennick 1982). PRPs account for approximately 70 % of all salivary proteins (Schenkels et al. 1995), making them an essential constituent of saliva substitutes for people suffering from xerostomia[13]. Unfortunately, these proteins can only be found in saliva, and in very small concentrations, making it very hard to purify them in the amounts necessary for academic and industrial purposes. Interestingly, a family of milk proteins called caseins have been shown to possess anticariogenic and remineralizing properties similar to those of PRP-1. β-casein (Ribadeau et al. 1972) is one of the most representative members of its family, constituting up to 45 % of all caseins (Farrell et al. 2004), and like PRP-1 it is also an IDP (Tompa 2002). Furthermore, while not a PRP *per se*, its sequence also contains a high number of proline residues. Thus, it has been hypothesized that β-casein could be used as a cheap and highly available alternative for PRP-1.

The absence of a well-defined structure in disordered proteins complicates the approach that must be taken when considering structural studies, since the most common goal, that is, the determination of a unique high-resolution structure, is not attainable for the isolated protein. Instead, the goal of such studies is usually to obtain information on the collection[14] of states that is sampled by the protein, including the estimation of its average size, shape and flexibility; and the detection of residual secondary structure, transient long-range contacts, and regions of restricted or enhanced mobility; with the hope that such information may prove informative regarding the associated biological function (Eliezer 2009). From an experimental point of view, this type of study can be quite challenging and thus the evergrowing importance of computer modeling and simulation of intrinsically disordered proteins (Rauscher & Pomès 2010).

---

[13] Xerostomia or "dry mouth syndrome" arises when there is a change in saliva composition, or if the saliva flow is reduced. There are various possible causes for this, including Sjögrens syndrome, diabetes, eating disorders, malnutrition and radiotherapy. It can be a side effect of various prescription drugs. Some of the most severe symptoms associated with xerostomia are tooth decay, tooth loss and an increased risk of infection (Napeñas et al. 2009).

[14] The most appropriate word here would be "ensemble", a concept which will be addressed further ahead in Section 2.1.

2 Theoretical background

In this section, a brief overview of the theoretical foundations required to understand how the properties of a system are connected to its microscopic behavior will be presented. As such, some of the most essential concepts of statistical thermodynamics and classic statistical mechanics need to be addressed. Special emphasis will be put into the theoretical description of the different types of intermolecular interactions, which are of paramount importance to the development of protein models.

2.1 Statistical thermodynamics

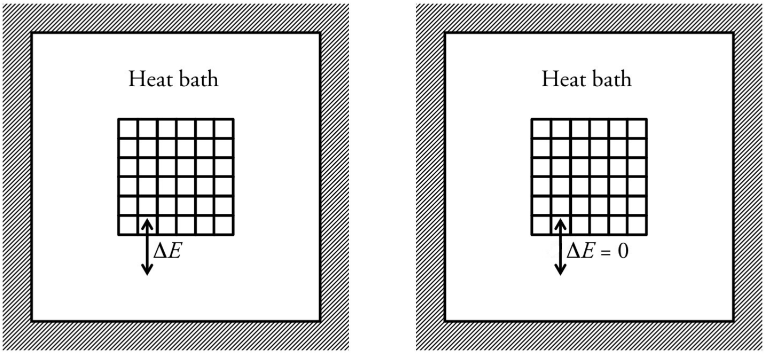
The object of statistical thermodynamics is to provide the molecular interpretation of equilibrium properties of macroscopic systems. In order to calculate macroscopic properties from molecular properties, it is necessary to set up postulates which allow us to proceed directly with this task as far as mechanical thermodynamic properties are concerned (e.g. pressure, energy, volume, number of molecules, *etc*). Non-mechanical properties (e.g. temperature, entropy, free energy, chemical potential, *etc*) are handled indirectly by classic thermodynamics. However, before presenting the postulates, we must first introduce the concept of an **ensemble** of systems.

An ensemble can be defined as a mental collection of a very large number $\mathfrak{N}$ of systems, each constructed to be a replica, on a thermodynamic level, of the reference thermodynamic system whose properties are being investigated. For a system of interest with volume V, containing N molecules, immersed in a large heat bath at temperature T, the assigned values of N, V and T are sufficient to determine its thermodynamic state. The respective ensemble would consist of $\mathfrak{N}$ systems, constructed to duplicate the thermodynamic state (N, V, T) and environment of the original system. All systems in the ensemble are identical from a thermodynamic point of view, but not on the molecular level, given that there is an extremely large number of quantum states consistent with the reference thermodynamic state. It is important to note that, due to the many different quantum states represented in the various systems of the ensemble, the calculated instantaneous value of any mechanical variable (not held constant) will differ depending on the quantum state. Hence, to obtain its average value, it is necessary to average over these instantaneous values, a procedure which is commonly referred to as "ensemble averaging". With this, we arrive at the **first postulate of statistical thermodynamics**:

> *The (long) time average of a mechanical variable $\mathcal{A}$ in the thermodynamic system of interest is equal to the ensemble average of $\mathcal{A}$, given that $\mathfrak{N} \to \infty$.*

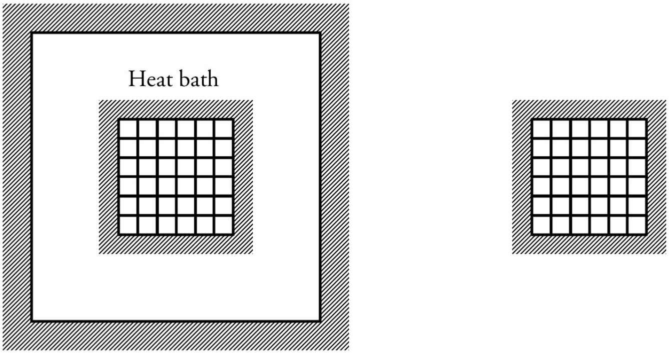
The **second postulate** states:

> *In an ensemble representative of an isolated thermodynamic system, the systems of the ensemble are distributed with equal probability over all possible quantum states.*



(**a**) A non-equilibrated canonical ensemble (with $\mathfrak{N}$ systems) exchanges energy with a very large heat bath.

(**b**) Energy equilibrium is achieved and the net energy transfer between the ensemble and the heat bath is zero.



(**c**) Thermal insulation can now be placed around the ensemble.

(**d**) The ensemble can be removed from the heat bath.

Figure 2.1: Thought experiment on how the canonical ensemble can be effectively treated as an isolated system, a condition which is necessary in order to apply the ergodic hypothesis. Notice that the ensemble in (d) is still considered canonical, given that each system is now in a "heat bath" composed of the remaining $\mathfrak{N} - 1$ systems.

The latter is usually called the "principle of *a priori* probabilities", and the word "isolated" is of key importance, since this postulate can be limited to the case of an isolated system, where N, V and the energy E are kept constant. The representative ensemble of such system is commonly called **microcanonical**. On the other hand, the first postulate is applicable to all different thermodynamical environments, of which the **canonical** (N, V, T) and **isothermal-isobaric** (N, p, T)[15] ensembles are the most relevant for this work. When the first and second postulates are combined, we arrive at the so-called **ergodic hypothesis**, which implies that a single isolated system spends equal amounts of time, over a long period of time, in each of the available quantum states. In other words, all accessible quantum states are equiprobable over a long period of time. The concept of ergodicity is of central importance for the statistical analysis of computational chemistry and biophysics.

---

[15] p stands for pressure.

As a general rule, most textbooks dedicated to this subject now make use of the these two postulates to derive the essential properties of the most commonly encountered ensembles, generally starting from the simplest case, i.e., the microcanonical ensemble. This is, however, out of the scope of this work, and I will instead focus on presenting the essential properties of the canonical ensemble, which is of greater relevance for the simulation of biophysical systems such as proteins in solution.

In the canonical ensemble, a given system has a fixed volume V, fixed number of molecules N, and it is immersed in a very large[16] heat bath at temperature T. *A priori*, because the thermodynamic system is not isolated but in contact with a heat bath, the energy of the system can fluctuate. However, once equilibrium is reached, thermal insulation can be placed around the outer boundaries of the ensemble, and the ensemble can be removed from the heat bath. In this procedure, the ensemble can be effectively treated as an isolated system (see Figure 2.1). Thus, the ergodic hypothesis holds and the basic statistical-mechanical equations that can be used to calculate the thermodynamic properties of a closed, isothermal system, can be derived.

As is customary, we should start by introducing the probability that the system is in any particular energy state E_j:

$$P_j(N, V, T) = \frac{e^{-E_j(N, V)/kT}}{Q(N, V, T)} , \tag{2.1}$$

where

$$Q(N, V, T) = \sum_j e^{-E_j(N, V)/kT} . \tag{2.2}$$

Here, Q is the so-called **partition function** - the canonical ensemble partition function, in this specific case - and it describes the equilibrium statistical properties of the system. Partition functions are functions of the thermodynamic state variables, and most of the aggregate thermodynamic variables of the system, such as the Helmholtz free energy A, entropy S, pressure p, total energy E, and chemical potential μ can be expressed in terms of the partition function or its derivatives:

---

[16] So that the limit $\mathfrak{N} \to \infty$ holds.

$$A(N, V, T) = -kT \ln Q(N, V, T) \;, \tag{2.3}$$

$$S = -\left(\frac{\partial A}{\partial T}\right)_{V,\,N} = kT\left(\frac{\partial \ln Q}{\partial T}\right)_{V,\,N} + k \ln Q \,, \tag{2.4}$$

$$p = -\left(\frac{\partial A}{\partial V}\right)_{T,\,N} = kT\left(\frac{\partial \ln Q}{\partial V}\right)_{T,\,N} , \tag{2.5}$$

$$E = -T^2\left(\frac{\partial A/T}{\partial T}\right)_{V,\,N} = kT^2\left(\frac{\partial \ln Q}{\partial T}\right)_{V,\,N} , \tag{2.6}$$

$$\mu = \left(\frac{\partial A}{\partial N}\right)_{T,\,V} = -kT\left(\frac{\partial \ln Q}{\partial N}\right)_{T,\,V} . \tag{2.7}$$

For a more comprehensive, yet accessible, presentation of the subjects discussed in this section, the reader is referred to Hill (1986, Chapters 1 and 2).

2.2 Classical statistical mechanics

The basics of statistical thermodynamics are commonly laid out from the quantum perspective, as it often provides a more general postulatory foundation. It is, however, worth noting that some of the most commonly used computer simulation methods for the study of the trajectories of atoms and molecules, make use of the laws of classical (Newtonian) mechanics for computing the motions within a system of interacting particles. Luckily, no results are obtainable from classical statistics which cannot be found as limiting laws from quantum statistics.

In the classical approach, the canonical partition function becomes:

$$Q_{\text{class}} = \frac{1}{N!\,h^{3N}} \int e^{-H(\mathbf{q},\ \mathbf{p})/kT} dx_1 \ldots dp_{zN} \,, \tag{2.8}$$

where h is Planck's constant, $H(\mathbf{q},\ \mathbf{p})$[17] is the Hamiltonian of a N components system with coordinates $\mathbf{q}$[18] and momenta $\mathbf{p}$[19]. This expression differs markedly from its quantum equivalent, as presented in Eq. 2.2. Firstly, the classical sum over all possible quantum states of the system is, in fact, an integral, since the classical state can vary continuously. Secondly, the classical energy is given by the Hamiltonian function, which is defined as:

$$H(\mathbf{q},\ \mathbf{p}) = \underbrace{\frac{1}{2m}(p_{x1}^2 + \ldots + p_{zN}^2)}_{\text{Kinetic energy}} + \underbrace{U(x_1, \ldots, z_N)}_{\text{Potential energy}} \;. \tag{2.9}$$

---

[17] Vectors are represented by upright boldface characters.

[18] $\mathbf{q} = x_1, \ldots, z_N$

[19] $\mathbf{p} = p_{x1}, \ldots, p_{zN}$

Integration of the kinetic term of the Hamiltonian can be carried out immediately, and we obtain a simplified form of Eq. 2.8

$$Q_{\text{class}} = \frac{Z_N}{N!\,\Lambda^{3N}}\,, \tag{2.10}$$

where

$$\Lambda = \frac{h}{(2\pi m\,kT)^{1/2}} \tag{2.11}$$

and

$$Z_N = \int_V e^{-U(x_1,\ldots,z_N)/kT} dx_1 \ldots dz_N\,. \tag{2.12}$$

Λ is the de Broglie wavelength and Z_N is called the classical **configuration integral**. The latter is of paramount importance for the calculation of the (canonical) ensemble average of a given property $\mathcal{A}$ (in the classical approach)[20]:

$$\begin{aligned} \langle \mathcal{A} \rangle &= \frac{\int_V \mathcal{A}(x_1,\ldots,z_N)\, e^{-U(x_1,\ldots,z_N)/kT} dx_1 \ldots dz_N}{\int_V e^{-U(x_1,\ldots,z_N)/kT} dx_1 \ldots dz_N} = \\ &= \frac{\int_V \mathcal{A}(x_1,\ldots,z_N)\, e^{-U(x_1,\ldots,z_N)/kT} dx_1 \ldots dz_N}{Z_N}\,. \end{aligned} \tag{2.13}$$

In-depth presentations of this subject, with examples of applications for simple systems such as the ideal gas, are given by Kjellander (2012, Chapter 4) and Hill (1986, Chapter 6).

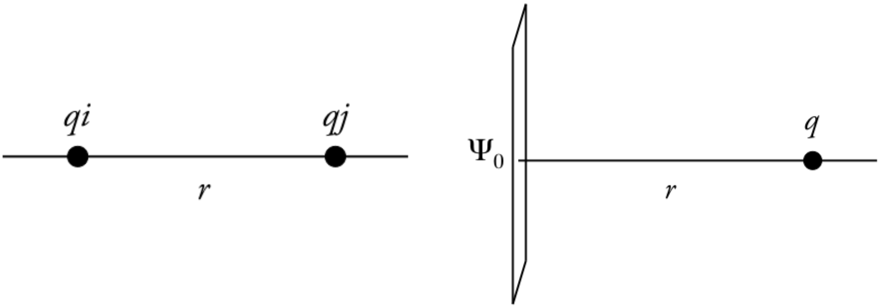
2.3 Intermolecular interactions

Intermolecular forces embrace all forms of matter and (bio)chemically relevant systems, such as a protein in solution, are no exception. According to Israelachvili (2011, Section 1.1), there are four distinct forces in nature. Strong and weak interactions, electromagnetic and gravitational interactions. The first two occur between subatomic particles and are the domain of nuclear and high-energy physics. The last one accounts for tidal motion and cosmological phenomena, and in most cases can be safely neglected when studying phenomena at the atomic scale. Electromagnetic forces, however, are usually the source of all intermolecular interactions, dictating the properties of solids, liquids, and gases, the behavior of particles in solution, chemical reactions, and the organization of biological structures. Electromagnetic interactions can generally be divided into two main groups:

Covalent – In this type of interaction, electron pairs are shared between atoms, giving rise to a chemical bond. It is thus intramolecular, complex and quantum mechanical

---

[20] Notice that $\mathcal{A} \neq A$. The latter usually stands for the Helmholtz free energy.



(**a**) Two charged particles.

(**b**) One charged particle and an ideal charged surface.

Figure 2.2: Two types of charge–charge interactions. (a) Two charged particles q_i and qj at a distance r, as in the Coulomb interaction (Eq. 2.16). (b) A charged particle q and a uniformly charged surface with surface charge potential Ψ_0, as in the Gouy-Chapman theory (Eq. 2.21).

by nature. In simpler, non-quantum mechanical simulations methods, the covalent bond is often approximated to a fixed length or as oscillating around an equilibrium distance, such as in a harmonic potential. This will be discussed in further detail in the next section.

Non-covalent – This class comprises the many different types of interactions or physical forces between non-bonded, discrete atoms and/or molecules. At least seven types of non-covalent or non-bonded intermolecular interactions can be present in biologically relevant systems, such as proteins in solution. All of these interaction types will presented in the following subsections. However, it should be noted that for most protein modeling purposes, it is common practice to limit the number of non-bonded interactions to be considered.

2.3.1 Charge–charge

As the name implies, charge–charge interactions describe the forces between two charged particles[21] and they are by far the strongest of the physical forces considered here. In certain cases it can even be stronger than most chemical binding forces.

The electric field **E** due to a point charge i a distance r away is:

$$\mathbf{E}_i = \frac{q_i}{4\pi\varepsilon_0\varepsilon r^2}\,, \tag{2.14}$$

where ε_0 and ε are the vacuum permittivity and the dielectric permittivity of the medium, respectively. The interaction of this field with a second charge q_j at r, gives rise to a force known as the **Coulomb force** or **Coulomb Law** (see Figure 2.2a):

---

[21] By "particles" I mean atoms and/or molecules, as the latter can be also be regarded as a point charge when represented in a simplified, coarse-grained formalism.

$$\mathbf{F}(r) = q_j\mathbf{E}_i = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r^2} \; . \tag{2.15}$$

The pair potential[22] for the Coulomb interaction between two charges q_i and q_j is therefore:

$$U(r)_{\text{Coulomb}} = \int_{\infty}^{r} -\mathbf{F}(r)\, dr = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r} \; . \tag{2.16}$$

Since $q_i = z_i\, e_c$, where z_i is the ionic valency of particle i and e_c is the elementary electron charge, we can further simplify Eq. 2.16 as:

$$\frac{U(r)_{\text{Coulomb}}}{kT} = \frac{e_c^2}{4\pi\varepsilon_0\varepsilon kT}\frac{z_i z_j}{r} \Leftrightarrow \beta U(r)_{\text{Coulomb}} = \lambda_B \frac{z_i z_j}{r} \; , \tag{2.17}$$

where λ_B is the so-called Bjerrum length and $\beta = 1/kT$.

In Eqs. 2.16 and 2.17, the only attenuation to the electrostatic interaction between charges q_i and q_j is due to the dielectric permittivity of the medium ε. However, it is often the case that the medium contains other ionic species, such as salt. The presence of these ionic species further enhances the aforementioned electrostatic screening between charges q_i and q_j. By making use of the Debye-Hückel theory, it is possible to arrive at a "screened" version of the Coulomb potential, which is given by:

$$\beta U(r)_{\text{Debye-Hückel}} = \lambda_B \frac{z_i z_j}{r}\, e^{-\kappa r} \; , \tag{2.18}$$

where κ is the inverse of the Debye (screening) length κ^{-1}, which can be written as:

$$\kappa^{-1} = \left(\frac{\varepsilon_0\varepsilon kT}{2N_A e_c^2 I}\right)^{1/2} \; . \tag{2.19}$$

In the previous equation, I is commonly referred to as the ionic strength and it is expressed as:

$$I = \frac{1}{2}\sum_k z_k^2 c_k \; , \tag{2.20}$$

with c_k being the concentration of ionic species k. It should be noted that the Debye-Hückel theory is approximate and works best for monovalent ions due to weak ion correlation effects[23]. Furthermore, the electrostatic screening is overestimated when

---

[22] Also referred to as free energy or available energy.

[23] To arrive at Eq. 2.18, a major approximation is used, that is, ion-ion correlation effects are entirely neglected. This means that we disregard the fact that ions of the same charge are repelled from the neighborhood of each other and ions of opposite charges are attracted to each other. It also means that we disregard the fact that ions cannot come closer to each other than the sum of their radii (Greberg et al. 1996). Luckily, for monovalent ions, this effect is not critical.

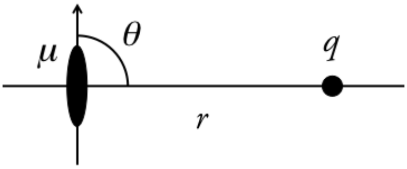


Figure 2.3: Schematic representation of the interaction between a charged particle q and a permanent dipole μ (Eq. 2.23). θ is the polar angle of the dipole.

the charges are in close contact, as a consequence of the continuum representation of salt ions.

Charge–charge intermolecular interactions are not restricted to "free" charged particles in solution. In fact, interactions occurring between charges fixed at a surface and those free in solution play an important role in many different research fields. For this particular work, the study of the interactions between proteins and charged surfaces, such as silica, are of great importance. The Gouy-Chapman theory relates surface charge density to surface potential and ion distribution outside a planar surface. In this theory, the surface is considered ideal, i.e., an infinite, uniformly charged planar surface exposed to an electrolyte solution, and the potential at any distance r from the surface is given by:

$$\beta\Psi(r)_{\text{Gouy-Chapman}} = \frac{2}{e_c} \ln\left(\frac{1+\Gamma_0\, e^{-\kappa r}}{1-\Gamma_0\, e^{-\kappa r}}\right) , \tag{2.21}$$

where

$$\Gamma_0 = \tanh\left(\frac{e_c\Psi_0}{4kT}\right) , \tag{2.22}$$

and Ψ_0 is the potential at the charged surface (see Figure 2.2b).

To arrive at the final expressions for the charge–charge interaction within the Debye-Hückel and Gouy-Chapman approaches (Eqs. 2.18 and 2.21, respectively), we must solve the Poisson-Boltzmann equation under different conditions, and the reader is referred to Israelachvili (2011, Chapter 14) and Evans & Wennerström (1999, Section 3.8) for more complete presentations of this subject.

2.3.2 Charge–dipole

Charge–dipole interactions occur between a charged particle and a polar molecule. For a charge q at a distance r from the center of a point dipole μ at an angle θ (see Figure 2.3), the corresponding pair potential can be written as:

$$U(r,\theta)_{\text{charge–dipole}} = -\frac{q\,\mu\cos\theta}{4\pi\varepsilon_0\varepsilon r^2} . \tag{2.23}$$

At large separations or in a medium of high dielectric permittivity ε, the angle dependence of this interaction falls below the thermal energy kT, and dipoles can rotate

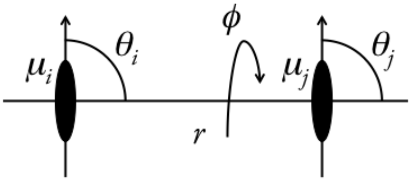


Figure 2.4: Schematic representation of the interaction between two permanent dipoles μ_i and μ_j (Eq.: 2.25). θ_i and θ_j are the respective polar angles, and ϕ is the azimuthal angle.

freely. Therefore, we can produce an angle-averaged charge–dipole pair potential:

$$U(r)_{\text{charge–dipole}} \approx -\frac{q^2\mu^2}{6(4\pi\varepsilon_0\varepsilon)^2 kTr^4} \quad \text{for} \quad kT > \frac{q\mu}{4\pi\varepsilon_0\varepsilon r^2} \ . \tag{2.24}$$

As can be seen, the interaction range decreases sharply, going from a $1/r^2 \rightarrow 1/r^4$ distance dependence.

For a practical demonstration of the derivation and approximations used in order to arrive at Eq. 2.24, the reader is referred to Israelachvili (2011, Section 4.10). The same rationale can be employed to obtain all other angle-averaged pair potentials presented below.

2.3.3 Dipole–dipole

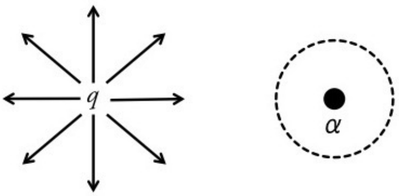
In the same manner that a charge can interact with another charge or a polar molecule (a permanent dipole), for two polar molecules near each other, with permanent dipoles μ_i and μ_j, the dipole–dipole interaction can be expressed by the following pair potential:

$$U(r,\theta_i,\theta_j,\phi)_{\text{dipole–dipole}} = -\frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon r^3}\left[2\cos\theta_i\cos\theta_j - \sin\theta_i\sin\theta_j\cos\phi\right] \ . \tag{2.25}$$

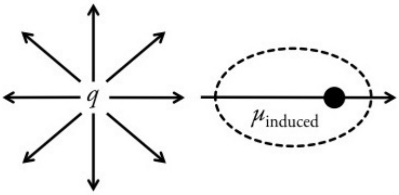
Notice that we now have to consider two polar angular dependencies θ_i and θ_j, and one azimuthal angle ϕ, in order to describe all rotational degrees of freedom of the two interacting dipoles (see Figure 2.4). However, as seen for the charge–dipole case, for weak interactions (relative to the magnitude of kT) it is possible to derive an angle-averaged dipole–dipole pair potential:

$$U(r)_{\text{Keesom}} = -\frac{\mu_i^2\mu_j^2}{3(4\pi\varepsilon_0\varepsilon)kTr^6} \quad \text{for} \quad kT > \frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon r^3} \ . \tag{2.26}$$

The angle-averaged dipole–dipole interaction between two permanent dipoles is one of the three fundamental interactions contributing to the **total van der Waals interaction** between atoms and molecules, and is commonly referred to as the **Keesom interaction.**



(**a**) A positively charged particle emits an electric field in the vicinity of a non-polar particle with polarizability α.



(**b**) The electric field from the positively charged particle attracts the electron cloud of the non-polar particle, resulting in its displacement relative to the nucleus, effectively inducing a dipole.

Figure 2.5: Schematic representation of the interaction between a charged particle q and a non-polar particle with polarizability α (Eq. 2.27). Here, the electron cloud is represented as a dashed circle/ellipse. The nucleus is represented as a filled black circle.

2.3.4 Charge–non-polar

All atoms and molecules are polarizable under the influence of an external electric field **E**. This is always true, regardless of whether these molecules were originally polar or non-polar by nature. This polarizability α arises from the displacement of the electron cloud relative to the positively charged nucleus, as a result of the applied field (see Figure 2.5). The polarization of an otherwise non-polar molecule is termed **induced dipole**. The charge–non-polar or, more appropriately, charge–induced dipole interaction pair potential is as follows:

$$U(r)_{\text{charge–induced dipole}} = -\frac{q^2\alpha}{2(4\pi\varepsilon_0\varepsilon)^2 r^4} \,. \tag{2.27}$$

For non-polar molecules, $\alpha = \alpha_0$. However, and as mentioned above, a permanent dipole can also be (further) polarized, and thus:

$$\alpha = \alpha_0 + \alpha_{\text{dipole}} = \alpha_0 + \frac{\mu^2}{3kT} \,, \tag{2.28}$$

making Eq. 2.27 of general application, that is, applicable to both non-polar and polar molecules.

2.3.5 Dipole–non-polar

By analogy with the previous example, a polar molecule is also able to interact with a non-polar molecule. The main difference is that in this case, the inducing field comes from a permanent dipole instead of a charge. Thus, for a fixed dipole μ oriented at an angle θ from a non-polar molecule with polarizability α_0:

$$U(r,\theta)_{\text{dipole–induced dipole}} = -\frac{\mu^2 \alpha_0 (1 + 3\cos^2\theta)}{2(4\pi\varepsilon_0\varepsilon)^2 r^6} \; . \tag{2.29}$$

The strength of this interaction is, however, commonly not strong enough to mutually align the molecules, as is the case for charge–dipole and dipole–dipole interactions. Therefore, the effective interaction is given by the angle-averaged energy:

$$U(r)_{\text{Debye}} = -\frac{\mu_i^2\, \alpha_{0_i} + \mu_j^2\, \alpha_{0_j}}{(4\pi\varepsilon_0\varepsilon)^2 r^6} \; . \tag{2.30}$$

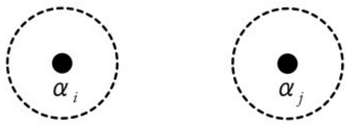
This is the second of the three inverse sixth power contributions to the total van der Waals interaction energy between molecules, and is often referred to as the **Debye interaction**.

2.3.6 Non-polar–non-polar

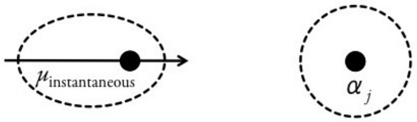
In addition to purely electrostatic interactions involving charged or dipolar molecules, there is another type of force, commonly known as **dispersion** or **London force**[24], that acts between *all* atoms and molecules. This force constitutes the third contribution to the total van der Waals interaction energy between molecules, and it is of paramount importance, due to the fact that it is *always present*, independently of the properties of the molecules being considered. Furthermore, this is the only possible interaction between two non-polar molecules. Dispersion forces are quantum mechanical by nature and their rigorous theoretical treatment is out of the scope of this work. In an overly simplistic manner, we can think of it as follows: even though the time average of the dipole moment of a non-polar atom (or molecule) is zero, at any given instant there exists a finite dipole moment due to the anisotropic distribution of its electrons around the nucleus. This instantaneous dipole generates an electric field that is capable of polarizing a nearby non-polar atom (or molecule), effectively inducing a dipole (see Figure 2.6).

---

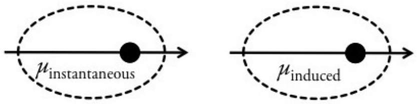
[24] Due to the major contribution of Fritz London to the study and understanding of the dispersion interaction.



(**a**) Two non-polar particles with polarizabilities α have $\langle \mu \rangle = 0$.



(**b**) However, their instantaneous $\mu \geq 0$.



(**c**) An instantaneous dipole of sufficient magnitude may induce a dipole on a second non-polar particle, leading to a stable instantaneous–induced dipolar interaction.

Figure 2.6: Schematic representation of the interaction between two non-polar particles with polarizabilities α_i and α_j. Note that for strictly non-polar molecules $\alpha = \alpha_0$, as in Eq. 2.31. The electron clouds are represented as dashed circles/ellipses, and the nuclei are represented as filled black circles.

London's result for the dispersion force is the following (London 1937):

$$U(r)_{\text{London}} = -\frac{3}{2} \frac{\alpha_{0_i} \alpha_{0_j}}{(4\pi\varepsilon_0)^2 r^6} \frac{h\nu_i \nu_j}{(\nu_i + \nu_j)} , \tag{2.31}$$

where h is Planck's constant, as encountered before, and ν is the orbiting frequency of the electron. For two identical atoms, the previous equation reduces to:

$$U(r)_{\text{London}} = -\frac{3h\nu\alpha_0^2}{4(4\pi\varepsilon_0)^2 r^6} . \tag{2.32}$$

One of the most obvious shortcomings of London's theory is that it cannot handle the interaction of molecules in a solvent, as is readily noticed by the absence of the solvent dielectric permittivity ε in Eqs. 2.31 and 2.32. McLachlan's theory (McLachlan 1965) is an alternative and more comprehensive approach, which covers the effect of the medium on dispersion forces in liquids.

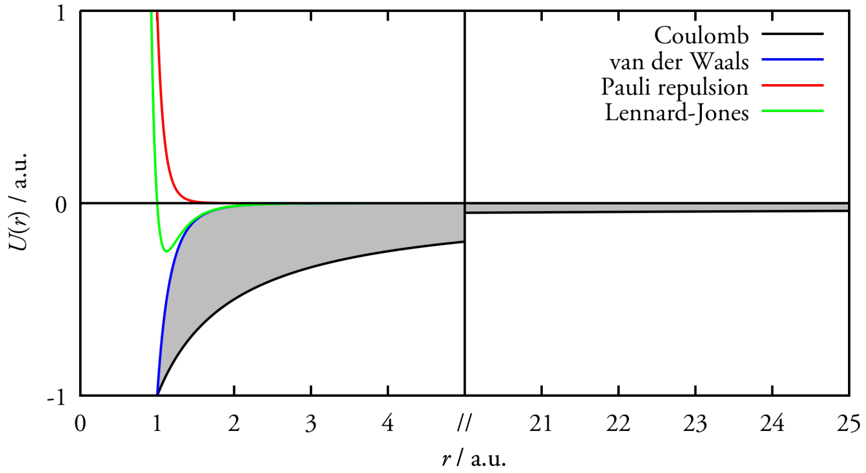


Figure 2.7: Graphical representation of the Coulomb and Lennard-Jones potentials (Eqs. 2.16 and 2.33, respectively), along with the individual contributions to the latter, that is, Pauli repulsion and inverse sixth power attractive distance dependence characteristic of van der Waals forces. The gray shaded area highlights the interaction strength and range difference between Coulombic and van der Waals forces.

2.3.7 van der Waals forces

As briefly mentioned throughout the previous subsections, the three inverse sixth power interaction potentials, namely the Keesom (Eq. 2.26), Debye (Eq. 2.30) and London (Eq. 2.31) interactions, comprise what is known as the **total van der Waals force**. However, it is common to see the term "van der Waals" employed in a rather loose manner, when referring to the latter of its constituents, that is, the London (or dispersion) interaction between strictly neutral, non-polar molecules. This is a result of the natural importance of this interaction, given its ubiquitous nature and often non-negligible contribution to the overall interaction energy, whenever many of such interactions are present in the system (which, more often than not, is the case).

That being said, the most commonly used mathematical expression used to model van der Waals interactions is the Lennard-Jones potential:

$$U(r)_{\text{Lennard-Jones}} = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] , \qquad (2.33)$$

where σ is the interparticle distance for which the potential is zero, and ϵ is depth of the potential. This potential is a composite, whose second term is responsible for the attraction between two particles, having the same inverse sixth power form as all the constituents of the total van der Waals force. This term alone is, however, not a good representation of reality, because maximum attraction will occur at $r = 0$ with an infinite strength and, in reality, two atoms or molecules cannot fully interpenetrate each other. Hence the need to include a term that takes into account the so-called **exchange interaction** or **Pauli repulsion**, which determines the closest interaction distance for

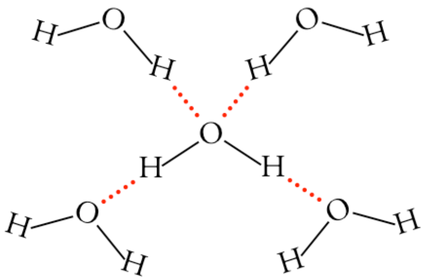


Figure 2.8: Example of a hydrogen bond network between water molecules. Even though four such interactions are possible, the exact number of hydrogen bonds formed by a molecule of liquid water fluctuates with time and depends on the temperature. The average number of hydrogen bonds per water molecule is around 3.59, for the TIP4P water model at 25 °C (Jorgensen & Madura 1985).

two particles. This repulsive term can be modeled either as "hard wall", i.e., $U(r) = \infty$ when r is below a certain threshold, or as "soft repulsion", by means of a $1/r^{12}$ type of potential.

Figure 2.7 shows the characteristic shape of the Lennard-Jones potential, along with its individual terms. The Coulomb potential (recall Eq. 2.16) is also shown to exemplify the difference in strength and effective interaction range between charge–charge interactions and van der Waals forces.

2.3.8 Hydrogen bond

The hydrogen bond can be described as a (mostly) electrostatic attraction between two polar molecules, occurring when a hydrogen atom covalently bonded to a highly electronegative atom such as oxygen, nitrogen or fluoride experiences the electrostatic field of another highly electronegative atom nearby. In simple terms, it can be considered a particularly strong case of the dipole–dipole interaction, but while the latter is usually not strong enough to lead to the mutual alignment of polar molecules in solution, the former can lead to fairly strong interactions with considerable directional character (see Figure 2.8). On a more rigorous theoretical ground, the nature of the hydrogen bond appears to be much more complex (Weinhold 1997), and opinions regarding whether it is purely electrostatic or if it also possesses some covalent character are divided in reference journal publications (Isaacs et al. 1999; Ghanty et al. 2000) and textbooks (Israelachvili 2011; Jackson 2006).

Apart from this controversy, it is widely recognized that hydrogen bonds play a key role in chemistry and biochemistry, e.g., determining the three-dimensional structures adopted by proteins and nucleic bases in DNA. Furthermore, the solvation of solute molecules by water and, consequently, the hydrophobic effect[25], are also in great part due to the hydrogen bond interaction.

---

[25] In simple terms, the hydrophobic effect can be described as the tendency of non-polar molecules to aggregate in aqueous solution. This is a complex phenomenon, which is in part due to the strong inclina-

Despite its relevance, there is no simple equation for the interaction potential that is satisfactorily predictive or accurate. The typical hydrogen bond strengths appears to fall under a $1/r^2$ distance dependence, but application of the charge–dipole interaction pair potential (recall Eq. 2.23) is not possible since the magnitude of the partial hydrogen charge is not known in advance.

---

tion of water molecules to form hydrogen bonds. By "clumping" all non-polar solute molecules together, the solute surface area is minimized and the number of water–water hydrogen bonds is maximized. This reduces the (considerable) energetic penalty arising from having to disrupt the bulk liquid structure in order to accommodate for the solute (Israelachvili 2011; Jackson 2006, Sections 8.5 and 2.8, respectively).

3 Modeling

In the previous section, the reader was presented with a brief overview of the foundations of statistical thermodynamics, its extension to the classical limit, and the different types of (intra- and) intermolecular interactions between atoms and molecules that, when taken together, make up the potential energy of the system. Obtaining a good model function for the potential energy which can weight the configurations of the system in a proper way is, in a way, the real secret behind the "art" of simulation modeling.

Since classical mechanics does not consider properties that depend upon the electronic distribution in a molecule, several valid assumptions have to be taken in account. The Born-Oppenheimer approximation, makes it possible to express the Hamiltonian of a system as a function of the nuclear variables (only), since the rapid motions of the electrons are averaged out. Therefore, the classical potential energy function, which is often called a **model**, tries to describe as accurately as possible the energy of the system using a rather simple model of the interactions within a system with contributions from processes such as the stretching of bonds, the opening and closing of angles, the rotations about single bonds (intramolecular forces); and non-bonded interactions (intermolecular forces), such as electrostatic and van der Waals interactions.

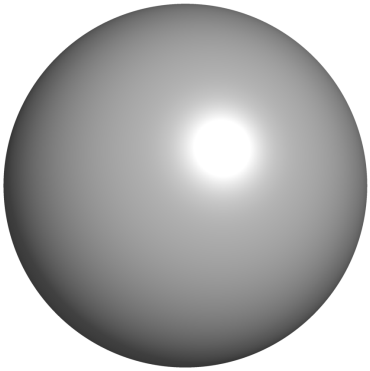
The degree of detail needed to preserve the dominant physical properties of a given system can vary considerably. Just as the Born-Oppenheimer approximation is often employed to reduce the complexity of a system for which electronic detail is not strictly necessary, even less detailed models - approximating molecules and building block residues, such as amino acids, to simple spheres (see Figure 3.1b) - can also produce very satisfactory results. This type of approximate formulation is commonly referred to as **coarse-graining** or **coarse-grained** modeling. **Atomistic** models (see Figure 3.1a), while also approximate when compared to quantum mechanical solutions, describe the system at the atomic level. In this work, both coarse-grained and atomistic models are used to describe systems of proteins in solution and in the presence of a charged surface. The motivation for using one model or the other is heavily correlated with the "scientific point of view" of each study, that is, what are the properties of interest under investigation and does the model capture the essential features of system accurately. Paraphrasing Einstein: "everything should be made as simple as possible, but no simpler."

3.1 Coarse-grained modeling of flexible proteins

In the coarse-grained simulations performed in this work, the intrinsically disordered proteins studied therein are modeled as flexible bead necklaces, where each amino acid residue and the termini are coarse-grained into soft, interpenetrating spheres, with or without point charges at their centers (depending on the nature of the given amino

(**a**) Fully atomistic representation of a protein.

(**b**) Coarse-grained representation of a protein, where each amino acid residue (and the terminal groups) are represented as spheres.



(**c**) Even more coarse-grained representation of a protein, where the entire structure is modeled as a single sphere.

Figure 3.1: Modeling of a penta-alanine peptide at different levels of detail. Despite the inherent differences between the models, it is possible that all of them capture (some of) the essential physics of the system, as required for different scientific endeavors.

acid residue). These charges are allowed to fluctuate during the simulation, in order to account for protein charge regulation. Furthermore, whenever it is of interest, an ideal surface with fixed charge density ρ is included in the simulation box. Solvent and salt effects are modeled implicitly through the Bjerrum and inverse Debye screening lengths (λ_B and κ, respectively; see Figure 3.2). The complete potential energy function for this coarse-grained model has the following form:

$$\begin{aligned} U(x_1, \ldots, z_N) = {} & U_{\text{bonds}} + U_{\text{Pauli repulsion}} + U_{\text{hydrophobic}} + U_{\text{Debye-Hückel}} \\ & + U_{\text{titration}} + \underbrace{U_{\text{Lennard-Jones}} + U_{\text{Gouy-Chapman}}}_{\text{For surface adsorption simulations only.}} \, . \end{aligned} \tag{3.1}$$

Notice that the last two terms are exclusive for surface adsorption simulations. All other terms are considered necessary in maintaining the essential features of (disordered) pro-

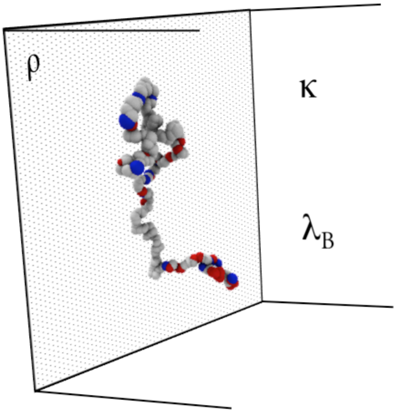


Figure 3.2: Illustration of a coarse-grained protein (β-casein, from Papers II and III) adsorbed onto a charged surface of charge density ρ. Solvent and salt are modeled implicitly, through the Bjerrum length λ_B and the inverse Debye screening length κ.

teins in solution. Further coarse graining would probably result in a non-representative model, with little to no application for the studies at hand, because model details are not only system specific but also tightly related to the goal of the study. The first five terms in Eq. 3.1 are expressed as:

$$U_{\text{bonds}} = \sum_{b} \frac{1}{2} k_b (r_{ij} - r_0)^2 \,, \tag{3.2}$$

$$U_{\text{Pauli repulsion}} = \sum_{i,j} 4\epsilon_r \left(\frac{\sigma_i + \sigma_j}{2\, r_{ij}} \right)^{12} \,, \tag{3.3}$$

$$U_{\text{hydrophobic}} = \sum_{i,j} \epsilon_h \;\; \text{for} \;\; r_{ij} \leq r_{cutoff} \,, \tag{3.4}$$

$$U_{\text{Debye-Hückel}} = \sum_{i,j} \lambda_B \, \frac{z_i z_j \, kT}{r_{ij}} \, e^{-\kappa r_{ij}} \,, \tag{3.5}$$

$$U_{\text{titration}} = \sum_{i} kT \ln 10 \, (\text{pH} - \text{p}K_{\text{a}\,i}) \;\; , \;\; \text{if } i \text{ is protonated} \,. \tag{3.6}$$

The bonded term (Eq. 3.2; b stands for bond) keeps residues i and j connected while allowing harmonic vibrations around the equilibrium bond distance r_0, and k_b is the respective force constant. As mentioned above, protein residues are modeled as soft spheres. The extent of this overlap is governed by a repulsive $1/r^{12}$ distance dependence (see Eq. 3.3), loosely referred to as Pauli repulsion, due to Wolfgang Pauli's exclusion principle[26]. In this term, ϵ_r is a parameter whose magnitude is inversely proportional

---

[26] Notice that Pauli's exclusion principle was not exactly formulated for this case, as it is quantum mechanical by nature, stating that two identical fermions cannot occupy the same quantum state simul-

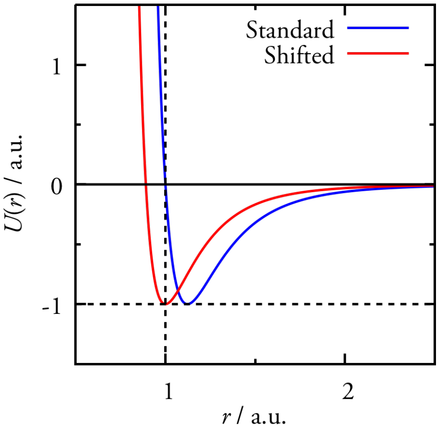


Figure 3.3: Difference between standard and shifted Lennard-Jones potentials (see Eqs. 2.33 and 3.7, respectively). For simplicity ϵ and σ are assumed to be unity. In the shifted version, maximum attraction occurs exactly at $r = \sigma$, while in the standard form it is verified at $r = 2^{1/6}\sigma \approx 1.222\sigma$.

to the degree of overlap allowed between beads, and σ_i and σ_j are the radii of beads i and j, respectively. In this model, two hydrophobic residues experience attraction to each other through a square well potential of magnitude ϵ_h (see Eq. 3.4), within a certain cutoff distance r_{cutoff}. Charged residues interact according to the Debye-Hückel theory (see Eq. 3.5), i.e., a "screened" version of the Coulomb potential as shown earlier (recall Section 2.3.1). Finally, if the deprotonated state of a titrable amino acid residue is chosen as the "reference state", with no overall contribution to the potential energy function, the contribution from the protonated state of the very same titrable residue takes the form shown in Eq. 3.6, where pK_{a_i} is the negative logarithm of the acid dissociation constant K_a for residue i.

As mentioned earlier, the last two terms in Eq. 3.1 are only applicable to simulations where an ideal charged surface is also present, and can be expressed as:

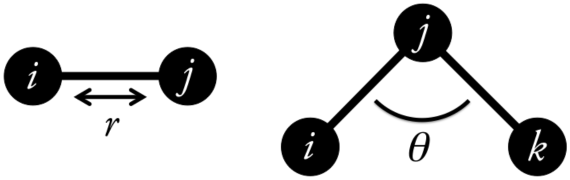
$$U_{\text{Lennard-Jones}} = \sum_i \epsilon_s \left[\left(\frac{\sigma_i}{r_{is}} \right)^{12} - 2 \left(\frac{\sigma_i}{r_{is}} \right)^{6} \right] , \tag{3.7}$$

$$U_{\text{Gouy-Chapman}} = \sum_i \frac{2kT}{e_c} \ln \left(\frac{1 + \Gamma_0 \, e^{-\kappa r_{is}}}{1 - \Gamma_0 \, e^{-\kappa r_{is}}} \right) . \tag{3.8}$$

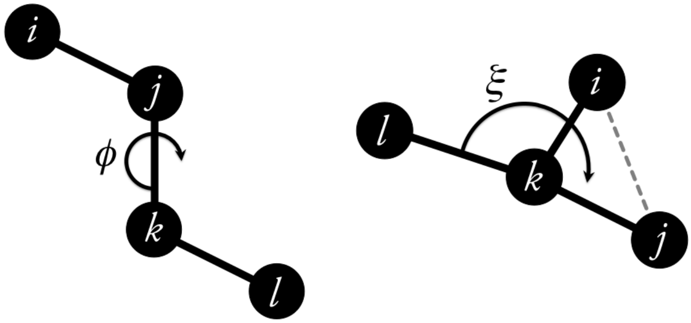
The first surface-specific term of the potential energy function is a shifted Lennard-Jones potential (see Eq. 3.7), which is used here to model non-electrostatic interactions (dispersion forces) between protein residues and the charged surface. ϵ_s is the depth

---

taneously. However, the main idea behind this principle is often carried into classical physics, as two particles cannot occupy the same exact position in space, at the same time.



(**a**) Bond (Eq. 3.10). (**b**) Angle (Eq. 3.11).



(**c**) Proper dihedral (Eq. 3.12). (**d**) Improper dihedral (Eq. 3.13).

Figure 3.4: Schematic representation of the different "bonded" interactions present in this atomistic potential energy function (Eq. 3.9).

of the potential, σ_i is the radius of particle i, and r_{is} is the distance between particle i and the surface s. As can be seen in Figure 3.3, the shifted Lennard-Jones potential has a more convenient form for this particular type of interaction, due to the fact that a particle will experience a maximum attraction towards the surface at close contact. With a standard Lennard-Jones potential (recall Eq. 2.33), the dispersion interaction between protein residues and the surface would be zero at the same distance, which is not a good representation of this interaction. Electrostatic interactions between charged residues and the charged surface are captured by using the Gouy-Chapman theory (see Eq. 3.8 and recall Section 2.3.1). Γ_0 is the same as in Eq. 2.22.

Finally, notice that (for simplicity) an informal summation notation is adopted throughout this work. Lower and upper bounds are omitted, and multiple summations are generalized into a single summation sign running over several indices, separated by commas.

3.2 Atomistic modeling of proteins

The atomistic simulations of IDPs, as presented in some of the publications included in this work, were performed with an atomistic model distributed in the Gromacs molecular dynamics simulation package (Hess et al. 2008; Pronk et al. 2013; Abraham et al.

2015). In this model, the following expression for the potential energy function is usually used to describe a variety of biomolecular systems, ranging from simple molecules to complex proteins, lipid bilayers and nucleic acids:

$$\begin{aligned} U(x_1, \ldots, z_N) = {} & \underbrace{U_{\text{bonds}} + U_{\text{angles}} + U_{\text{proper dihedrals}} + U_{\text{improper dihedrals}}}_{\text{bonded interactions}} \\ & + \underbrace{U_{\text{Coulomb}} + U_{\text{van der Waals}}}_{\text{non-bonded interactions}} \ . \end{aligned} \tag{3.9}$$

Here, the first two terms are related to the harmonic constraints in the bond (two-body) and angle (three-body) values, respectively. The third and fourth terms are related to the four-body dihedral angle torsions. Altogether, the following equations represent the bonded interactions related with the covalently bonded atoms (van der Spoel et al. 2014):

$$U_{\text{bonds}} = U(r_{ij}) = \sum_b \frac{1}{2} k_b (r_{ij} - r_0)^2 \ , \tag{3.10}$$

$$U_{\text{angles}} = U(\theta_{ijk}) = \sum_\theta \frac{1}{2} k_\theta (\theta_{ijk} - \theta_0)^2 \ , \tag{3.11}$$

$$U_{\text{proper dihedrals}} = U(\phi_{ijkl}) = \sum_\phi k_\phi \left[1 + \cos(n\phi_{ijkl} - \delta)\right] \ , \tag{3.12}$$

$$U_{\text{improper dihedrals}} = U(\xi_{ijkl}) = \sum_\xi k_\xi (\xi_{ijkl} - \xi_0) \ . \tag{3.13}$$

Eqs. 3.10 and 3.11 represent, respectively, the harmonic vibrations of bonds r_{ij}, around the equilibrium bond length r_0; and bond angles θ_{ijk}, around the equilibrium angle θ_0 (see Figures 3.4a and 3.4b). k_b and k_θ represent the respective force constants. The aforementioned operators are often regarded as "hard" degrees of freedom, due to the substantial energies that are required to cause significant deformations from their reference values. Most variation in structure and relative energies is due to the complex interplay between the torsional and non-bonded contributions. Torsional potentials represent the ability (or inability) of a bond to rotate around its own longitudinal axis. It is inherent to adjacent four-body dihedral angles *ijkl*. Proper dihedral angles (Eq. 3.12) are defined according to the IUPAC/IUB[27] convention, where ϕ_{ijkl} is the angle between the *ijk* and *jkl* planes, with zero corresponding to the *cis* configuration (see Figure 3.4c). The periodic behavior of these interactions can be described by a sinusoidal function with periodicity n and phase δ, and k_ϕ establishes the height of the torsion energetic barrier. Some dihedrals, called improper dihedrals, are however meant to keep planar groups planar (e.g. aromatic rings) or to prevent molecules from flipping over to their mirror images. They are commonly written in the form of a harmonic term (see Eq. 3.13) that treats out-of-plane distortions and maintain chirality (see Figure 3.4d). The potential energy due to an improper dihedral ξ depends on the equilibrium dihedral ξ_0, and the force constant k_ξ. The summation indices b, θ, ϕ and ξ stand for any pair, triplet or quadruplet of atoms that form a bond, angle, proper or improper dihedral, respectively.

---

[27] International Union of Pure and Applied Chemistry / International Union of Biochemistry.

For the non-bonded interactions in Eq. 3.9, U_{Coulomb} is the Coulomb interaction between two charges, as presented in Eq. 2.16, but here one must now account for all pairs of interactions between charged particles in the system:

$$U_{\text{Coulomb}} = U(r_{ij}) = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \; . \tag{3.14}$$

The non-polar–non-polar component of the inverse sixth power van der Waals forces is, once more, approximated by the use of a Lennard-Jones potential:

$$U_{\text{van der Waals}} = U(r_{ij}) = \sum_{i,j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \; , \tag{3.15}$$

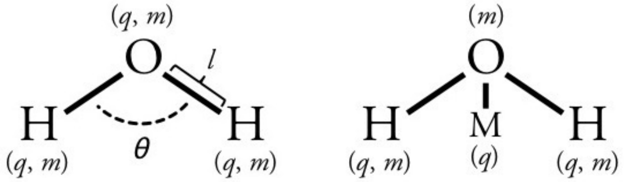
where ϵ_{ij} is the depth of the potential well, and σ_{ij} is the finite distance at which the interparticle potential is zero. The Lennard-Jones potential can be written in alternative ways, and for this specific formulation ϵ_{ij} and σ_{ij} are calculated based on the Lorentz-Berthelot combination rules:

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\, \epsilon_{jj}} \; , \tag{3.16}$$

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj}) \; . \tag{3.17}$$

3.3 Explicit water models

The two previous subsections deal with computer modeling of proteins at different levels of detail. However, apart from the coarse-grained model case, where the solvent is treated implicitly, no mention has been made regarding how water, the solvent in all the studies presented this work, can be explicitly modeled. As it happens, the appropriate treatment of solute–solvent and solvent–solvent interactions is a key factor for the outcome of most atomistic simulations, yet, it is common to see most of the effort in protein modeling and force field development being directed towards the protein itself. Intrinsically disordered proteins are specially sensitive to the choice of the water model, as they are significantly exposed to the solvent, due to the (often) extended conformational ensembles they tend to adopt on solution (Best et al. 2014; Palazzesi et al. 2014; Piana et al. 2015; Henriques et al. 2015; Rauscher et al. 2015; Ye et al. 2015; Mercadante et al. 2015; Henriques & Skepö 2016).



(**a**) 3-site water model. (**b**) 4-site water model.

Figure 3.5: Schematic representations of the structures of (a) 3-site and (b) 4-site water models. The 3-site water model contains one interaction site per atom. Each atom has a partial charge q and a mass m. l is the O–H bond length, and θ is the H–O–H angle. SPC and TIP3P are both 3-site water models. The 4-site water model contains a dummy site M for the charge of the oxygen atom. The delocalization of the oxygen partial charge improves the electrostatic distribution around the water molecule. TIP4P/2005s and TIP4P-D are examples of 4-site water models. There are more complex water models, where an additional number of dummy sites are used to represent the lone pairs of the oxygen atom, for example. Note: l and θ are not represented in (b) for schematic simplicity.

There is a multitude of water models which can be used to simulate water molecules explicitly. For brevity, only the water models which were used in the publications included in this thesis will be presented, that is, the simple point charge (**SPC**) (Berendsen et al. 1981), **TIP3P** (Jorgensen et al. 1983), **TIP4P/2005s** (Abascal & Vega 2005; Best et al. 2014) and the **TIP4P-D** (Piana et al. 2015) water models. All these models are non-polarizable, i.e., the partial charges and dipole moments are conserved throughout the simulation, regardless of the external electric fields they may eventually be exposed to[28], and they can be categorized according to the respective number of interaction sites they possess (see Figure 3.5). The interaction of explicit water molecules among themselves and with the solute (and explicit salt ions, when applicable) is mediated by the same potential energy function as presented before (recall Section 3.2). Bonded and non-bonded interactions follow the same exact formalism, and the main difference among water models is mainly related to how many interactions sites are present in their construction and, as equal or more importantly, their unique parameters for the equilibrium bond length, angle, partial charges and van der Waals interaction strength (see Table 2). The rationale behind the derivation of such parameters is largely responsible for the effect that water will have, as a solvent, on other biomolecules.

As a final comment, it should be noted that despite the inclusion of explicit water molecules in a simulation, the dielectric permittivity of the medium (which is also water) is still considered whenever calculating the Coulomb force between *any* two charged particles in the system, as all interactions are considered pairwise additive, and the effect of all the other water molecules is approximated by a single averaged effect. This is mainly due to the fact that many-body interactions are often impractical or even impossible to solve, and we must settle for approximations.

---

[28] As is also the case for the protein models presented before.

Table 2: List of parameters for the water models used in this work. q, l and θ are the partial charge, bond length and angle, respectively, as shown in Figure 3.5. σ and ϵ are the typical Lennard-Jones parameters, as encountered several times before (recall Section 2.3.7, for example). For TIP4P/2005s, two ϵ values are tabulated. The value signaled with * is used for the calculation of water–water and water–salt interactions, and † is specific to water–protein interactions. σ and ϵ are zero for water hydrogen atoms, which means that these atoms do not contribute to the van der Waals interaction term of the potential energy function (recall Eq. 3.9).

parameter	units	water model			
		SPC	TIP3P	TIP4P/2005s	TIP4P-D
$q(\text{O})$	e_c	0.82	0.834	0	0
$q(\text{H})$	e_c	0.41	0.417	0.556	0.58
$q(\text{M})$	e_c	-	-	1.113	-1.16
l	Å	1	0.957	0.957	0.957
θ	degrees	109.47	104.52	104.52	104.52
$\sigma(\text{O})$	Å	0.317	0.315	0.316	0.317
$\epsilon(\text{O})$	kJ mol^{-1}	0.651	0.636	0.775^{*} / $0.938^{\dagger}$	0.937

3.4 Considerations about force fields

Models and force fields are necessarily interconnected, and it is often the case that they are used interchangeably when referring to the functional form and parameter set used to calculate the potential energy of a system of atoms and molecules (or just particles, as in a coarse-grained approach). For the specific purpose of this work, it seems more appropriate to make an explicit distinction between them. Here, a model reflects the physical model that governs all the interactions in the system, and it takes the form of a potential energy function (recall Eqs 3.1 and 3.9), with individual terms for specific types of bonded and non-bonded interactions. On the other hand, in order to define a force field, we must not only specify the functional form, but also its intrinsic parameters. Variables like bond length r_{ij}, bond angle θ_{ijk}, and dihedral angle ϕ_{ijkl}, among others, can be determined from the positions of each particle in the system. Yet, the various constants for each potential term, such as the bond force constant k_b and the reference bond length r_0, are specific to each force field.

The quality of the force field is heavily determined by the quality of its parameters. Hence, force field parameterization is of extreme importance for bringing simulations into a fully quantitative and accurate level. However, regardless of the parameterization undergone for a certain force field, it will never be able to reproduce all properties of a system. In fact, a force field will generally predict certain properties better than others. Transferability of the functional form and its parameters is thus an important feature of a force field. Ideally, we would like to use the same set of parameters to model a wide range of molecules of interest. Yet, since force field parameters are generally obtained from quantum mechanical calculations and/or by fitting experimental data for a finite (and often related) set of compounds, it is not possible to obtain general parameters that satisfy all molecular structures. Therefore, most force fields are designed to handle a series of related molecules. Some force fields can be designed to handle a wider range of molecular systems, but the price for their generality comes in the form of a hefty decrease in accuracy, which makes their use limited to rough initial guesses of how a certain system may behave.

As an example, let us consider the case of the atomistic simulation of structured and intrinsically disordered proteins, as presented in Paper IV. The same molecular dynamics force fields and water models that have been traditionally used to simulate folded proteins with considerable success, perform very poorly when employed in the simulation of IDPs. Yet, at the most basic level, folded and disordered proteins can appear virtually indistinguishable, as, by definition, single-chain proteins are just a sequence of amino acid residues joined by peptide bonds[29]. Thus, *a priori*, it may not be entirely obvious why the aforementioned force fields and water models are not able to produce good results for the simulation of IDPs. This shows that we must be very careful when selecting a given model and set of parameters for the simulation of a specific system. Preliminary testing is key and, for this specific example, we should not assume that a set of parameters originally derived for folded proteins should be directly transferable to unfolded and disordered proteins.

---

[29] Note that this is a gross oversimplification. The full picture is clearly much more complicated than this, because folded and disordered proteins are enriched in either order- and disorder-promoting amino acid residues, respectively. This leads to completely different behaviors in aqueous solution.

4 Monte Carlo simulations

The multidimensional integral over particle coordinates, that is, the classical configuration integral (see Eq. 2.12), can only be analytically computed for a few exceptional cases. In all other cases, approximations or numerical techniques must be used to be able to compute the ensemble average of a given property (see Eq. 2.13). The Monte Carlo method is one of such techniques. The simplest Monte Carlo technique is denominated **random sampling** or **brute force** Monte Carlo. It works in the following manner: Say we are interested in obtaining the average value of a given function $f(x)$. In this approach, the unweighted average $\langle f(x) \rangle$ is determined by evaluating $f(x)$ at a large number $\mathfrak{N}$ of x values randomly distributed over the phase space. At $\mathfrak{N} \to \infty$, this procedure should yield the correct value. However, while conceptually easy to understand, this method is of little use to evaluate ensemble averages due to the fact that most of the computational effort is spent on points where the Boltzmann factor is negligible. It would instead be much more efficient to sample most points in the regions of space that make important contributions to the integral. This concept is called **importance sampling**.

4.1 The Metropolis method

The **Metropolis method** or algorithm, is one of the most recognized importance sampling techniques for Monte Carlo simulations. It tackles the problem of solving Eq. 2.13 from a different perspective. Instead of focusing on the configuration integral, we can instead try to determine the ratio of the two integrals. Metropolis et al. (1953) showed that it is possible to devise an efficient Monte Carlo scheme to sample such a ratio. We start by noticing that the ratio $e^{-U(x_1,\ldots,z_N)/kT}/Z_N$ in Eq. 2.13 is the probability density of finding the system in a given configuration, that is:

$$P(x_1, \ldots, z_N) \equiv \frac{e^{-U(x_1,\ldots,z_N)/kT}}{Z_N} \,. \tag{4.1}$$

This means that, if we were somehow able to randomly generate points in the coordinate space according to this probability distribution, on average, the number of points n_i generated per unit volume is equal to $LP(x_1, \ldots, z_N)$, where L is the total number of points generated, i.e.:

$$\langle \mathcal{A} \rangle \approx \frac{1}{L} \sum_i n_i \, \mathcal{A}(\mathbf{q}_i) \,. \tag{4.2}$$

To generate points in configuration space with a relative probability proportional to the Boltzmann factor, we start by defining an initial configuration $\mathbf{q}$, hereupon denoted as o (old), having a non-vanishing Boltzmann factor $e^{-U(\mathrm{o})/kT}$. A new trial configuration $\mathbf{q}'$, from here on denoted as n (new), is then generated by adding a small random displacement δ to o. The Boltzmann factor of this trial configuration is $e^{-U(\mathrm{n})/kT}$. In equilibrium, the transition probability from o to any available n states must be equal to the transition probability from any of these n states back to o. This detailed balance conditions means that:

$$P(\mathrm{o})\,\pi(\mathrm{o} \to \mathrm{n}) = P(\mathrm{n})\,\pi(\mathrm{n} \to \mathrm{o}) \;, \tag{4.3}$$

where $\pi(\mathrm{o} \to \mathrm{n})$ is the transition probability from an old configuration to a new one. Thus, the probability of accepting a trial move from o to n is:

$$P_{\mathrm{acc}}(\mathrm{o} \to \mathrm{n}) = \frac{\pi(\mathrm{o} \to \mathrm{n})}{\pi(\mathrm{n} \to \mathrm{o})} = \frac{P(\mathrm{n})}{P(\mathrm{o})} = e^{-[U(\mathrm{n})-U(\mathrm{o})]/kT} \,. \tag{4.4}$$

Because the acceptance probability cannot exceed unity, we have to consider the following:

$$P_{\mathrm{acc}}(\mathrm{o} \to \mathrm{n}) = \begin{cases} e^{-[U(\mathrm{n})-U(\mathrm{o})]/kT} & \text{if } U(\mathrm{n}) > U(\mathrm{o}) \\ 1 & \text{if } U(\mathrm{n}) \leq U(\mathrm{o}) \end{cases} \tag{4.5}$$

Clearly, whenever $\delta U \leq 0$, the acceptance probability is unity and the trial move must be accepted. However, to decide whether to accept or reject a trial move for which $\delta U > 0$, a random number $\mathfrak{r}$ is generated from a uniform distribution in the interval $[0, 1]$, and the following criterion is applied:

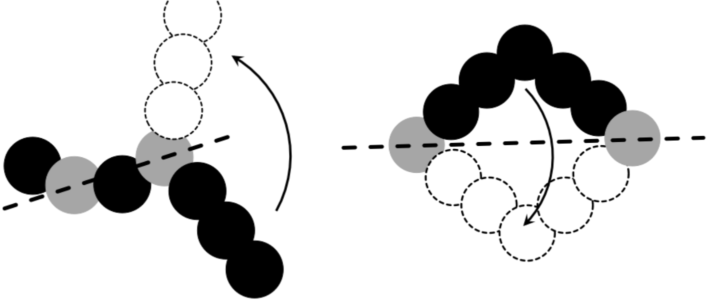
- $\mathfrak{r} < e^{-[U(\mathrm{n})-U(\mathrm{o})]/kT}$, the trial move is accepted;
- $\mathfrak{r} \geq e^{-[U(\mathrm{n})-U(\mathrm{o})]/kT}$, the trial move is rejected.

The Metropolis algorithm is iterative, repeating the procedure described above until convergence is achieved, generating a Markov chain. Moreover, by following these rules, we guarantee the sampling of points in the coordinate space with probability proportional to the Boltzmann factor, consistent with the theory of equilibrium statistical mechanics (thus rendering the method ergodic).

To end, we can then compute average properties by summing them along the path followed through all sampled configurations (recall Eq. 4.2).

4.2 Trial moves

Unlike molecular dynamics simulations, where Newton's equations of motion define the trajectories for all atoms in the system (as will be seen in Section 5), there is no strict "recipe" for generating new trial moves in Monte Carlo simulations. Moreover, even completely unreasonable and "unphysical" types of moves will eventually lead to a proper sampling of the system's coordinate space, given a long enough number of iterations. However, this will most likely lead to a high trial move rejection rate, which is exactly the opposite of what is usually aimed for, as - for obvious reasons - we are



(**a**) Pivot move. (**b**) Crank shaft move.

Figure 4.1: Schematic illustration of the (a) pivot and (b) crank shaft trial moves. The gray shaded circles represent the two randomly select particles from which the rotational axis (black dotted line) is defined. Dotted circles represent the new positions of the moving particles.

generally interested in choosing the most efficient sampling procedure for the system at hand.

Intrinsically disordered proteins are flexible by nature and can thus adopt many different conformations. The following five types of trial moves were found appropriate for this type of system (Evers et al. 2012; Cragnell et al. 2016; Hyltegren et al. 2016), and were thus employed in the studies of IDPs in solution and/or in the presence of a charged surface (Papers I, II and III):

Atomic translation – A single particle of a group is selected for a spatial translation within the simulation box. A displacement parameter sets the maximum magnitude of the translation. This parameter is determined by considering the simulation box size, the density of the system and the strength of the interactions. An optimal displacement parameter maximizes the root mean square displacement during the simulation, while asserting efficient sampling, i.e., maintaining a high trial move acceptance ratio.

Group translation and rotation – Similar to the atomic translation move, but instead of focusing on a single particle, this move translates and rotates an entire group as a whole. This move is of no consequence for a single protein in a box with both implicit solvent and salt, given that the change in potential energy is always zero, as there are no other interacting groups in the system. Therefore, within the Metropolis algorithm, this trial move would always be accepted, providing no real advantages at the expense of computational iterations and simulation time. On the other hand, this trial move is of paramount importance for the proper sampling of systems consisting of multiple explicit components, such as several interacting proteins, protein(s) in explicit salt, protein(s) in the vicinity of a charged surface, *etc*.

Pivot – This move and the next are specific to polymer-like molecules. Here, two residues within a polymer chain are randomly selected and the vector that connects them is used as the rotational axis for all the particles that sit at one of its ends (see Figure 4.1a).

Crank shaft – Similar to the pivot move, but here the residues in between the two randomly selected residues are rotated instead (see Figure 4.1b).

Titration – In order to study the charge regulation of a protein in solution and/or in the presence of a charged surface, a so-called titration or charge swap move is necessary. To allow charge fluctuations during the simulation, a trial charge swap is applied to a randomly picked ionizable residue with an acceptance probability:

$$P_{\text{acc}} = \min\left[1, e^{-[\delta U \pm \ln 10(\text{pH}-\text{p}K_{\text{a}})]/kT}\right] , \tag{4.6}$$

where $+$ is applied for protonation and $-$ for deprotonation.

4.3 Example of a basic algorithm

Basic Metropolis Monte Carlo algorithm

Here is an example of a Monte Carlo algorithm for a system consisting of a single protein in solution, using the Metropolis criteria for trial move acceptance or rejection:

1. Generate a random initial configuration (and protonation states) by placing all protein residues at random location within the simulation box.
2. Calculate the interaction energy $U(\text{o})$, based on the set of coordinates (and protonation states).
3. Generate a random integer between $[1, M]$, where M is the total number of trial move types.
4. Select the trial move type according to the outcome of the last step.
5. Execute the trial move.
6. Calculate the new interaction energy $U(\text{n})$, and:
    - (a) Accept the trial move if $U(\text{n}) \leq U(\text{o})$.
    - (b) Otherwise, generate a random number $\mathfrak{r}$ between $[0, 1]$ and accept trial move if $\mathfrak{r} < e^{-\delta U/kT}$.
    - (c) If neither points no. 7 or 8 apply, reject the trial move and restore the previous state.
7. Go back to point no. 2 and repeat.

Points no. 2 to 7 constitute what is called a simulation loop. This main loop can be arranged into a series of nested loops. It is common to have at least two loop levels (a macro and a micro loop), such that a simulation with say, 10^7 total iterations, can be "divided" into N_{macro} and M_{micro} loops, such that $N_{\text{macro}} \times M_{\text{micro}} = 10^7$ iterations. The main advantage of this procedure is that computationally expensive routines, such as the calculation of the radius of gyration and other system properties, do not have to be computed at every single iteration. These routines can instead be computed once, every macro loop iteration, greatly reducing the number of sampling events, and significantly increasing the performance of the simulation with no detrimental effect on the accuracy of the final averages[30].

---

[30] As long as that property is properly converged.

5 Molecular dynamics simulations

Molecular dynamics solves Newton's equations of motion for a given molecular system, and ultimately generates the trajectories for all atoms in the system. Therefore, it provides a way to calculate the microscopic interactions of the system, generating a representative ensemble of configurations, which will be essential to the computation of its macroscopic behavior.

5.1 Equations of motion

There are several different techniques that can be employed to solve the classical equations of motion for a system of N molecules interacting through a potential $U(x_1, \ldots, z_N)$ (see Eq. 3.9). Here, the Lagrangian equation of motion will be used, since it is considered to be the most fundamental form to describe motion (Allen & Tildesley 1987, Section 3.1):

$$\frac{\partial \mathcal{L}}{\partial q_k} = \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k}\right) . \tag{5.1}$$

The Lagrangian function $\mathcal{L}(\mathbf{q}, \dot{\mathbf{q}})$ is defined in terms of kinetic and potential energies:

$$\mathcal{L} = K - U , \tag{5.2}$$

and is a function of the generalized coordinates q_k and their time derivatives $\dot{q}_k$. **Newton's second law of motion** states that:

> *A body of mass m, subject to a force* $\mathbf{F}$*, undergoes an acceleration* $\mathbf{a}$*, that has the same direction as the force and a magnitude that is directly proportional to the force and inversely proportional to the mass, i.e.,* $\mathbf{F} = m\,\mathbf{a}$.

Then, if we consider a system of atoms, with Cartesian coordinates $\mathbf{r}_i$ and the usual definitions of K and U (see Eqs. 2.9 and 3.9), Eq. 5.1 becomes:

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i , \tag{5.3}$$

where m_i is the mass and $\ddot{\mathbf{r}}_i$ is the second time derivative of the Cartesian coordinates (i.e., the acceleration) of atom i. The force $\mathbf{F}_i$ acting on each particle in the system can be determined by the gradient of the potential energy U, relatively to the position of each atom i:

$$\mathbf{F}_i = \nabla_{\mathbf{r}_i} \mathcal{L} = -\nabla_{\mathbf{r}_i} U = -\frac{\partial U}{\partial \mathbf{r}_i} . \tag{5.4}$$

The gradient is a function of all the atomic coordinates $\mathbf{q}$, of the N particles that constitute the system at a given time.

Since forces are vectorial quantities and the potential energy is a scalar quantity, it is only natural that in molecular dynamics, the forces are calculated as the negative derivatives of all the analytic expressions describing the potential energy function.

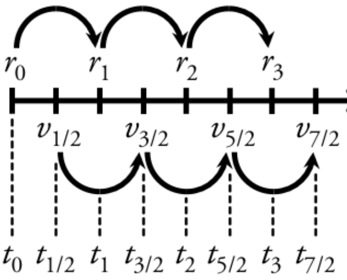


Figure 5.1: Schematic representation of the leap-frog algorithm. Calculated positions r and velocities v "leapfrog" over each other at each integration time step t, as in Eqs. 5.5 and 5.6. Velocities at integer time steps are then are calculated according to Eq. 5.7.

Once the force acting on all atoms is calculated we can integrate Newton's equation of motion and obtain the particle's new positions and velocities. However, this can only be accomplished by using numerical methods, which will be discussed next.

For a comprehensive derivation of the equations of motion from the Lagrangian formulation of classical mechanics see Frenkel & Smit (2002, Appendix A).

5.2 Finite difference methods

The equations of motion are solved assuming that the potential energy of the system U, is a continuous function of particle positions. Since the use of a continuous potential implies that the motions of all particles are coupled together, the equations of motion described in the previous section become a many-body problem, which is impossible to solve analytically. Therefore, these equations are integrated using a finite difference method.

There are several algorithms available for integrating the equations of motion using finite difference methods. In a general way, these algorithms assume that the positions and dynamic properties of a system can be approximated as Taylor series expansions, in which the integration is done iteratively at a fixed time interval δt. The total force on each particle at a time t is calculated as the vector sum of its interactions with other particles. Once the force is determined, the accelerations of the particles are calculated. The combination of the accelerations with the positions and velocities at time t (which are known from the last iteration) enables the calculation of the new positions and velocities at a time $t + \delta t$. This protocol is repeated for each and everyone of the following time steps. However, there is one important assumption that needs to be noted: the force is regarded as being constant during the time step of each iteration.

In this work, the leap-frog algorithm (Hockney & Eastwood 1988, Section 4.6) was used for the integration of the equations of motion. Therefore, the following discussion will be dedicated to it, and, for brevity, no other finite difference methods will be addressed. This algorithm is derived from the basic Verlet scheme (which will not be presented here), but overcomes Verlet's algorithm main deficiencies: its awkward handling of the velocities (since this term is not present explicitly in the formulation), and the inherent needless introduction of numerical imprecision through the introduction of a small term to a difference of large terms, while generating the trajectory.

At a first step, the velocities $\mathbf{v}$ at time $t + \frac{1}{2}\delta t$ are evaluated from both the velocities at time $t - \frac{1}{2}\delta t$ and the accelerations $\mathbf{a}$ at time t:

$$\mathbf{v}\left(t + \frac{1}{2}\delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\delta t\right) + \delta t\,\mathbf{a}(t)\ . \tag{5.5}$$

The positions $\mathbf{r}$ are then updated for a time $t + \delta t$, based on the velocities calculated before together with the positions at time t:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t\,\mathbf{v}\left(t + \frac{1}{2}\delta t\right)\ . \tag{5.6}$$

Finally, it is possible to determine the velocities, at time t, from the following expression:

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}\left(t + \frac{1}{2}\delta t\right) + \mathbf{v}\left(t - \frac{1}{2}\delta t\right)\right] \tag{5.7}$$

This last step is necessary so that the energy - as defined in Eq. 2.9 - at time t can be determined, which requires positions and velocities computed at the same instant.

As can be see from Eqs. 5.5 and 5.6, the velocities leap over the coordinates to give the next mid-step values at $t + \frac{1}{2}\delta t$. After this is done, the positions leap over the velocities, yielding their new values at $t + \delta t$. The first step is performed again, giving a new set of velocities at $t + \frac{3}{2}\delta t$, and so on (see Figure 5.1). Thus the name "leap-frog".

For additional information about different finite difference methods, the reader is referred to Allen & Tildesley (1987, Section 3.2) and Frenkel & Smit (2002, Sections 4.2.3 and 4.3.1).

6 Simulation techniques

In Sections 4 and 5, the fundamentals of both Monte Carlo and molecular dynamics simulations were formally presented. They constitute the kernel of the aforementioned simulation methods. However, additional simulation techniques are needed in order to run an efficient simulation. If we think of a simulation as a car, it can safely be said that the engine is probably the most important part, but without chassis and wheels, it is going nowhere. Since Monte Carlo and molecular dynamics programs share a number of structural features, i.e., both start from an initial configuration of molecules, which is then updated to a new set of configurations in a particular ensemble, ending up with the calculation of observable properties by averaging over a finite number of iterations; most of the ideas presented in this section are applicable to both simulation methods. Particular exceptions will be noted explicitly.

6.1 Periodic boundary conditions

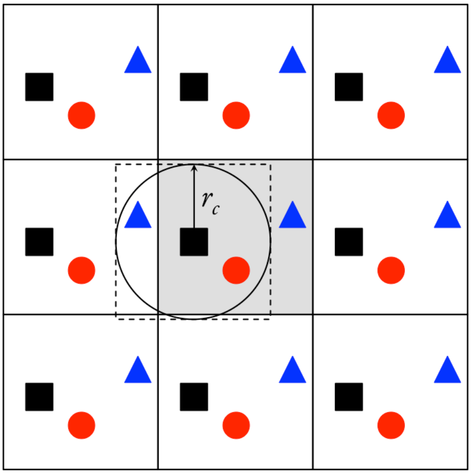
To simulate a "realistic" system consisting of a protein solution, the simulation box would need to contain a great number of proteins and an even greater number of water molecules and ions (for atomistic models). However, such a system would be too complex and too expensive to simulate with current computational resources. Furthermore, even if it was computationally feasible to simulate such system, the percentage of molecules in contact with the simulation box boundaries would be very large, e.g. $43\,\%$ for a cubic crystal of 10^6 atoms[31] (Frenkel & Smit 2002, Section 3.2.2). Thus, to simulate bulk phases adequately, it is essential to choose boundary conditions that mimic the presence of an infinite bulk surrounding the (central) model system. This is achieved by employing **periodic boundary conditions**.

Periodic boundary conditions imply the replication of the solution containing vessel in all directions, yielding a periodic array (see Figure 6.1a). If the mentioned vessel is, for example, a cubic box (since it is easier to visualize), this means that it will be surrounded by images of itself throughout space to form an infinite lattice. The images and the central box behave in a completely identical fashion. Whenever a molecule leaves the central box, one of its images will enter through the opposite face (as exemplified by the cartoon in Figure 6.1b). There are no walls at the boundary of the central box, and no surface molecules. In fact, the box is not intended to serve as container of the solution, it is just a convenient axis system for measuring the (internal) coordinates of the system's N molecules. Even though the cubic box is the simplest periodic system to visualize and to program, maximum performance is generally achieved by using system-dependent simulation box geometries. In fact, the cubic cell is potentially one of the least desirable geometries[32], specially for simulations containing explicit water

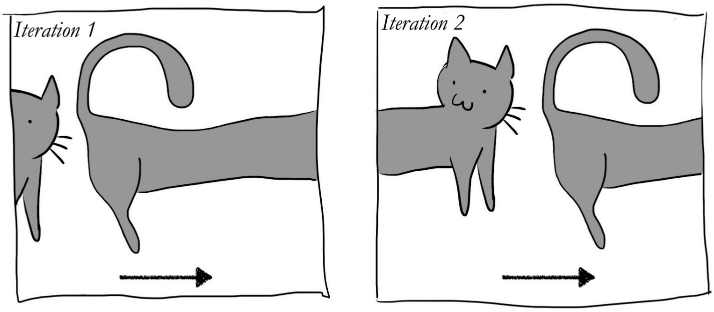
---

[31] This number increases even further as the size of the system is decreased.

[32] Since folded proteins are often globular and IDPs generally act like random coils in solution.



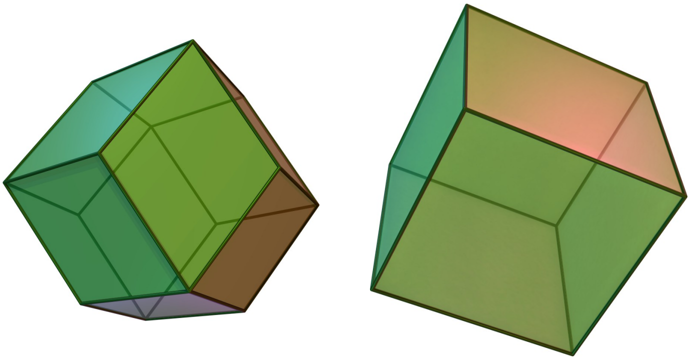
(a) Schematic representation of a 2-dimensional system and its periodic images (only the first shell is shown).



(b) Cartoon illustration depicting translation on a periodic system. Adapted from Martín (2012).

Figure 6.1: Schematic representation of periodic boundary conditions. In (a), according to the minimum image convention, a particle (central black square) may interact another particle in the neighboring cell (blue triangle) as long as it is closer than the equivalent particle in its own simulation cell (gray shaded square). The interaction with the latter is thus ignored.

molecules, as a considerable number of these molecules, present near the cube vertices, are generally too far away from the solute to be able to interact appreciably. Still, they must be explicitly accounted for in the potential energy calculations. Hence, it is sensible to choose a periodic cell that reflects the underlying geometry of the system. Take for example, the rhombic dodecahedron (see Figure 6.2a). It is one of the smallest and most regular space-filling unit cells for which periodic boundary conditions may be applied. Its volume is 71 % of the volume of a cube having the same image distance (see Figure 6.2).


(a) Rhombic dodecahedron, taken from Wikipedia (2016c).

(b) Cube, taken from Wikipedia (2016a).

Figure 6.2: Side-by-side comparison of the shapes and sizes of (a) a cube and (b) a rhombic dodecahedron.

6.2 Potential truncation and the minimum image convention

The calculation of the potential energy of a system subject to periodic boundary conditions implies several considerations. For a system of N particles it is not possible to calculate the force on particle i, or those contributions to the potential energy involving particle i, assuming pairwise additivity. This is because on top of having to include all the interactions between particle i and the remaining $N - 1$ particles in the original simulation box, we would also need to include the infinitely many interactions with its images. This leads to an infinite sum, which is obviously impossible to calculate in practice. However, it is often the case that in system consisting of, for example, a protein in solution, most relevant interactions are relatively short-range[33], and we may restrict this summation by making an approximation, i.e., it is possible to truncate the potential by applying the **minimum image convention**.

The minimum image convention states that each particle should see, at most, just one image of every other particle in the system. Thus, energy (and force) calculations are made considering only the closest particles or particle images. This is possible through the establishment of a spherical cutoff r_c, which must not exceed half of the shortest box vector. The interactions between all pairs of particles that are further apart than the cutoff value are not considered (see Figure 6.1a).

As suggested above, the application of the minimum image convention represents an approximation to the calculation of the "real" potential energy of the system. However, it is important to reiterate that this approximation is considered sensible in most cases, as a large contribution to the potential energy (and forces) comes from neighboring particles that are often closer to the reference particle than r_c. In particular, for a typical coarse-grained simulation of a single protein, where the solvent and salt are treated implicitly, it is feasible to effectively truncate all terms of the potential energy function (recall Eq. 3.1), because the size of the simulation box can be increased significantly[34] with little to no effect on the overall performance of the simulation. However, for atomistic simulations containing explicit solvent and salt molecules, simulation performance is a serious issue and the application of the minimum image convention is generally limited to short-range non-bonded potentials such as the van der Waals term in Eq. 3.9. Long-range potentials, such as the Coulomb interaction (Eq. 3.14), usually display an interaction range greater than half the box length for a system of moderate size (recall Figure 2.7), and need to be handled in a more elaborate fashion.

---

[33] Due to the shielding effect of the solvent (and salt).

6.3 Long-range force handling

As mentioned in the previous subsection, van der Waals interactions decay rapidly with the increasing distance between two interacting particles (recall Figure 2.7). This decay is usually complete within half the simulation box length, and thus the minimum image convention is applicable. Yet, whenever treating charge–charge interactions, this is often not true. Therefore, methods for handling long-range forces are of significant importance. Two of the most relevant are the **generalized reaction field** (Tironi et al. 1995) and the **particle-mesh Ewald** (Darden et al. 1993) methods, with the latter being the current *de facto* method in most molecular dynamics simulation packages. A short description of their main features will be made here.

Generalized reaction field – The reaction field method assumes that electrostatic interactions of a reference particle i with other particles beyond a certain cutoff distance can be handled in an average way, i.e., using macroscopic electrostatics, while the short-range contribution arising from interactions with particles situated within the same cutoff sphere is explicitly considered in the calculations. The particles outside the spherical cutoff are considered to form a dielectric continuum, producing a reaction field within the inner sphere or cavity. The *generalized* reaction field method is a development of the original reaction field, in which the dielectric continuum beyond the cutoff also has an ionic strength contribution. This method has the advantage of being conceptually simple, easily implemented and efficient. Furthermore, the possibility of introducing the ionic strength as an external parameter makes it even more attractive. It does, however, suffer from two specific problems. First, there is a discontinuity in the energy when the number of particles within the cutoff sphere of particle i changes. This results in poor energy conservation. Second, one needs to know the external dielectric constant beforehand.

Particle-mesh Ewald – The particle-mesh Ewald method is an improvement of the original Ewald summation method, which was first introduced as a method to calculate long-range interactions of the periodic images in crystals, and is now commonly used for calculating long-range interactions in computational chemistry. The basic idea behind the Ewald summation is that we can think of long-range interactions as having two major contributions, i.e., (i) a short-range contribution, and (ii) a long-range contribution which does not have a singularity. While the former can be easily handled in real space, the latter is calculated in reciprocal space using a Fourier transform. The advantage of this method is the rapid convergence of the energy compared with that of a direct summation. The main drawback is that the computational cost of the reciprocal part of the sum increases as N^2, which makes it impractical for large systems. This issue is, however, solved by the particle-mesh Ewald method, in which the reciprocal space sum is approximated by a multidimensional interpolation, inspired by the particle-mesh method of Hockney & Eastwood (1988, Section 1.5.2). The approximate reciprocal energy and forces are expressed as convolutions and can thus be evaluated quickly using fast Fourier transforms. The resulting algorithm scales as $N \ln N$, and is thus substantially faster than ordinary Ewald summation.

---

[34] To a point where r_c can safely be made larger than any possible interaction range.

To end, it should be mentioned that the particle-mesh Ewald method is not restricted to charge–charge interactions, and, if desired, can be used with van der Waals interactions as well.

For more comprehensive presentations on how to handle long-range forces, the reader is referred to Allen & Tildesley (1987, Section 5.5) and Frenkel & Smit (2002, Chapter 12).

6.4 Neighbor lists

After having shown how to enhance the performance of computer simulations through the truncation of short-range potentials and how to handle long-range forces, it is now appropriate to introduce the concept of a **neighbor list**. By knowing which particles to include in the non-bonded calculations within the established cutoff(s), it is possible to avoid computationally intensive tasks such as looping over all $N - 1$ particles, determining minimum images, calculating distances and checking if they are within the cutoff. Since in the simulation of systems in the condensed phase the neighbors of a given particle do not change significantly over a few iterations, it is possible to employ a method for determining neighbor particles lying within the cutoff range. The resulting "neighbor list" is updated every so many simulation steps, and is used to differentiate which particles are to be included or not in the non-bonded calculations. Several different neighbor list methods exist, but the main principle is the same as mentioned above, i.e., saving CPU time by reducing the frequency of computationally intensive tasks that are not strictly necessary at every simulation step.

Allen & Tildesley (1987, Section 5.3) and Frenkel & Smit (2002, Appendix F) provide an in-depth overview of this subject.

6.5 Bond and angle constraints

The main objective behind the establishment of bond and angle constraints, in molecular dynamics, is to enable bigger integration time steps without losing important conformational information. To achieve this, we have to institute a balanced compromise between what motions and interactions can be treated in an approximate manner and which need to be taken into account in their explicit form. For example, torsional motions are of lower frequency than bond vibrations, and the conformational information that can be obtained from torsional motion analysis is of much higher importance than that given by bond vibrations. Therefore, the iterative integration of the equations of motion at say, 2 fs, is generally a good compromise. This is because most bonds vibrate with a frequency above this value. However, valuable torsional motions occur at a lower frequency and are therefore explicitly followed with a 2 fs time step. Without constraints, the time step in molecular dynamics simulation would be dictated by the highest frequency motion present in the system, i.e., bond vibrations. This way, integration time steps would be too short and relevant simulation times (in the order of micro- to millisecond) would be very difficult to achieve with today's computational resources.

Several methods for constraining bonds and angles in molecular dynamics exist, but their individual presentation and discussion is inconsequential for this work. It should, however, be noted that the words *constraint* and *restraint* should not be used interchangeably. A constraint is a requirement that the system is forced to satisfy, that is, a constrained bond is forced to adopt a specific value throughout the entirety of the simulation. On the other hand, by applying a restraint on a bond, we are simply encouraging the system to adopt this value. There is no attempt to force it to adopt the value set as a restraint. In fact, the system is free to deviate from the optimal value, but it will incur in a (considerable) energetic penalty.

6.6 Constant temperature and pressure

In the molecular dynamics method, the total energy of the system is a constant of motion, and if we assume that the time averages are equivalent to ensemble averages, then the time averages obtained in a conventional simulation are representative of the ensemble averages in the microcanonical ensemble (N, V, E). However, this is often not the most convenient ensemble, as to mimic biologically relevant systems it would be more appropriate to use the canonical (N, V, T) or isothermal-isobaric (N, p, T) ensembles.

There are several ways to control the temperature of the system. The simplest implies scaling the velocities, achieved through the multiplication of the velocities at each time step by a scaling factor λ. The **Berendsen temperature coupling** method (Berendsen 1991), maintains the temperature by coupling the system to an external heat bath that is fixed at the desired temperature. The scaling of the velocities is done such that the rate of change of temperature is proportional to the difference in the temperature between the bath and the system. The bath, thus, acts as a source of thermal energy, supplying or removing heat from the system as needed. The following expression shows how the temperature changes with regard to time:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} , \tag{6.1}$$

where τ is the coupling parameter that determines how tightly the bath and the system are coupled together. The scaling factor λ modifies the velocities of each particle every n_{TC} steps, and is expressed in the following manner:

$$\lambda = \left[1 + \frac{n_{\mathrm{TC}}\,\delta t}{\tau_T} \left(\frac{T_0}{T\left(t - \frac{1}{2}\delta t\right)} - 1 \right) \right]^{1/2} . \tag{6.2}$$

The parameter τ_T is related to the time constant τ of the temperature coupling as follows:

$$\tau = \frac{2 C_v \tau_T}{N_{df} k} , \tag{6.3}$$

with C_v being the heat capacity of the system and N_{df} is the total number of degrees of freedom.

The Berendsen thermostat has one major drawback, that is, it suppresses the fluctuations of the kinetic energy. This means that it does not generate a proper canonical ensemble and, rigorously, the sampling will be incorrect. The **velocity-rescaling thermostat** (Bussi et al. 2007) corrects this issue by introducing an additional stochastic term that ensures a correct kinetic energy distribution:

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{K K_0}{N_{df}}}\frac{dW}{\sqrt{\tau_T}} . \tag{6.4}$$

Here, K is the kinetic energy and dW is a stochastic process called Wiener process. No additional parameters are necessary and this thermostat produces a correct canonical ensemble.

Most methods used to control the pressure are similar to those used for temperature control. A scaling factor is once again present in the rationale behind these techniques, and the pressure can be maintained at a constant value by simply scaling the volume of the simulation box. By analogy with the Berendsen thermostat, a Berendsen barostat also exists (Berendsen et al. 1984), but, once again, it does not yield the exact isothermal-isobaric ensemble. Thus, a different approach is often recommended. The **Parrinello-Rahman** pressure coupling (Parrinello & Rahman 1981; Nosé & Klein 1983) gives the true isothermal-isobaric ensemble by the use of an extended Lagrangian, i.e., the Lagrangian (recall Eq. 5.2) is extended to contain additional, artificial terms, which modify the Lagrangian equations of motion (Eq. 5.1) in such a way that the box volume is considered as a dynamical variable.

7 Simulation analyses

In this section, some of the most recurring types of analyses performed in the publications included in this thesis will be addressed.

7.1 Size, shape and stiffness

Since flexible, disordered proteins do not exhibit a stable 3-dimensional structure in solution, fine structural analyses are often overlooked in favor of the calculation of more coarse properties such as the average size, shape and stiffness. When taken together, these properties should provide a good indication of which IDP conformational[35] class (Das et al. 2015) the disordered protein being analyzed belongs to. Furthermore, these three properties are also obtainable from experimental methods, such as small-angle X-ray scattering, and being able to perform a direct comparison between the calculated values and experimental reference is of paramount importance for simulation validation.

The size of an IDP is generally obtained through the calculation of the **radius of gyration**:

$$R_g = \left(\frac{\sum_i \|\mathbf{r}_i\|^2 m_i}{\sum_i m_i} \right)^{1/2} , \tag{7.1}$$

where $\|\mathbf{r}_i\|$ is the (Euclidean) distance between the position $\mathbf{r}_i$ of atom/particle i and the molecule's center of mass, and m_i is the mass of i. In simple terms, this property describes how, on average, the components of a protein are distributed around its center of mass.

The shape of a flexible polymer chain can either be obtained by normalizing the end-to-end distance with the contour length[36], or by determining the so-called shape factor, i.e., the ratio between the end-to-end distance and the radius of gyration. The **end-to-end distance**, at any given instant, can be easily determined by the following expression:

$$R_{\text{end-to-end}} = \sqrt{\|\mathbf{r}_{\text{NT}} - \mathbf{r}_{\text{CT}}\|^2} , \tag{7.2}$$

where $\mathbf{r}_{\text{NT}}$ and $\mathbf{r}_{\text{CT}}$ are the positions of the N- and C-termini. For coarse-grained models, these positions are taken from the center of mass of the terminal beads, and for atomistic models these positions correspond to the α carbon of the terminal residues.

---

[35] Notice the use of the word "conformational" instead of "configurational". Until now, the latter has been used rather loosely, but in the context of a chain molecule, i.e., where specific building blocks (e.g. amino acid residues in a protein) appear in a specific order or sequence, a different "configuration" implies a different ordering of these elements, i.e., a different sequence. However, what is actually meant here is that the atoms of a protein can adopt a collection of different spatial arrangements, as a result of, e.g., rotations about individual bonds, angles and dihedrals, without changing the protein's amino acid sequence. These spatial arrangements are, by definition, called "conformations".

[36] The length of the polymer at its maximum physically possible extension.

Regardless of its size and shape, the stiffness or rigidity of a protein can be always be roughly estimated by the magnitude of the variance (and standard deviation) associated with the distribution of the sampled radii of gyration and the end-to-end distances. In other words, the stiffness of a protein chain should be inversely proportional to how much its average size and shape fluctuates (assuming equilibrium), and flexible proteins should present a significant dispersion around the average radius of gyration $\langle R_g \rangle$, and average end-to-end distance $\langle R_{\text{end-to-end}} \rangle$. A stricter measure of stiffness can be borrowed from polymer theory, where the **persistence length**[37] is the mechanical property usually employed to quantify the stiffness of a freely jointed polymer chain. Proteins, whether disordered or not, are heteropolymers of amino acids, and thus the calculation of the persistence length should apply. We must, however, be aware that the solutions in polymer theory often involve a number of approximations, and polymers are, for example, often assumed to be infinitely long, which is far from true in the case of some of the IDPs studied in the publications included in this work.

For comprehensive presentations about the structural properties of polymers/proteins (in solution), in connection with polymer theory, the reader is referred to Evans & Wennerström (1999, Section 7.1) and Jackson (2006, Chapter 3).

7.2 Total charge and charge capacitance

The net charge Z of a protein is easily determined by summing up all the (partial) charges of the constituting atoms/particles:

$$Z = \sum_i z_i \,, \tag{7.3}$$

where z_i is the (partial) charge number of atom/particle i. For constant-pH simulations, that is, simulations where titrable residues are allowed to protonate and deprotonate according to the selected solution pH, the protein net charge will fluctuate around its average value $\langle Z \rangle$. The variance of the protein net charge is:

$$\mathrm{var}(Z) = \langle Z^2 \rangle - \langle Z \rangle^2 \,. \tag{7.4}$$

It can be shown (see Lund & Jönsson (2013) for the mathematical proof) that by exposing the protein to an external potential Ψ_{ext}, the response in $\langle Z \rangle$ is:

$$\frac{\partial \langle Z \rangle}{\partial \Psi_{\text{ext}}} = \langle Z^2 \rangle - \langle Z \rangle^2 = C \,. \tag{7.5}$$

Here, C is a capacitance, more specifically, the **charge capacitance**. C is an intrinsic property of the protein, closely related to protein structure and sequence. It is also a function of the solution conditions, such as pH and ionic strength. Moreover, the charge capacitance can be used to estimate how an external potential, such as that from a charged surface, influences the protonation state of a protein. The higher the charge capacitance of a given protein, the higher is its ability to regulate its own (net) charge, which often leads to rather interesting physical mechanisms.

---

[37] The persistence length is defined as the length over which two parts of the chain keep their orientational correlation.

As is immediately noticeable from Eqs. 7.4 and 7.5, $C \equiv \mathrm{var}(Z)$, which means that the calculation of the charge capacitance of a protein simulated with a constant-pH method is rather trivial. Additionally, C can also be easily derived from the experimental titration curve of a protein by the following relation (Lund & Jönsson 2013):

$$C = -\ln 10 \, \frac{\partial Z}{\partial \mathrm{pH}} \, . \tag{7.6}$$

7.3 Principal component analysis

The complete energy landscape of a molecule is a function of all conformational coordinates, and it contains all the information required to build physically meaningful conformation classes. However, the complete specification of the conformation of a system of N atoms requires $3N-6$ internal coordinates, which is an intractable number of dimensions, even for relatively small systems. Luckily, in most cases we are merely interested in characterizing some sort of low-dimensional energy landscape that captures the relevant behavior of the system as a function of a small set of coordinates that represent the system in a simple way. To reduce the complete ($3N-6$)-dimensional conformational space onto a low-dimensional representation that is able to retain the most important features of the distribution of conformations, we can take advantage of the **principal component analysis** (PCA) method.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, it accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors form an uncorrelated orthogonal basis set.

Although the number of principal components obtained from PCA are as many as the original coordinates, the general idea is that we should keep only the amount needed to reasonably capture the conformational distribution. Since the first two or three principal components typically hold about 70–90 % of the total variance[38], we typically display this density as a function of the first two principal components. For a protein system, it is also common practice to analyze the backbone atoms only, thus removing several dimensions of the complete landscape. Protein translation and rotation can also be partially removed by fixing its center of mass and performing a least-squares fitting of the trajectory to a reference structure. Due to the absence of experimental reference structures for disordered proteins, it seems reasonable to use the the central structure of the simulation, that is, the conformation i, among the N sampled conformations, which minimizes the following dispersion measure (Campos & Baptista 2009):

---

[38] This number is system-dependent.

$$D_i = \left(\frac{1}{N-1} \sum_{i,j} \mathrm{RMSD}_{ij}^2 \right)^{1/2} , \tag{7.7}$$

where RMSD is the root mean square deviation between conformations i and j[39]. In non-mathematical terms, the central structure of a simulation is the protein conformation which differs the least from all other sampled conformations.

In order to construct a 3-dimensional energy landscape based on the two first principal components, a (bivariate) kernel density estimator is used to define the probability density function $P(\mathfrak{r})$ at all points $\mathfrak{r}$[40], which can then be transformed into a conditional free energy (or potential of mean force) - denoted here as E - through the following relation:

$$E(\mathbf{r}) = -kT \ln \frac{P(\mathfrak{r})}{P_{max}} . \tag{7.8}$$

Here, P_{max} is the maximum value of $P(\mathfrak{r})$. Through this "normalization" a zero energy is assigned to the maximum of the probability density. The resulting energy landscape can not only be of aid in visualizing and comparing the conformational landscapes of different simulations, but can also be further analyzed in order to determine which groups of conformations belong to which minima, i.e., it can prove invaluable in identifying the distinct conformational classes of a protein (see Figure 7.1).

7.4 Small-angle X-ray scattering

Small-angle X-ray scattering (SAXS) is an experimental technique that allows the determination of the dimensions, shape and flexibility of a protein in solution. Due to the inherent difficulty in determining the structure-function specificities of intrinsically disordered proteins using classical structural methods, such as X-ray diffraction or nuclear magnetic resonance (NMR), SAXS is becoming increasingly valuable. Not only because it is effective, but also due to the fact that it is particularly well adapted to the study of such proteins, being one of the few techniques that can characterize the unfolded/disordered state of proteins (Receveur-Bréchot & Durand 2012).

---

[39] $\mathrm{RMSD}(t_1, t_2) = \left[\frac{1}{\sum_k m_k} \sum_k m_k \|\mathbf{r}_k(t_1) - \mathbf{r}_k(t_2)\|^2 \right]^{1/2}$, for two different conformations obtained at simulation times t_1 and t_2. $\mathbf{r}_k$ and m_k are the position and mass of atom k, respectively.

[40] Here, $\mathfrak{r}$ denotes the vector containing all values of the principal components being considered.

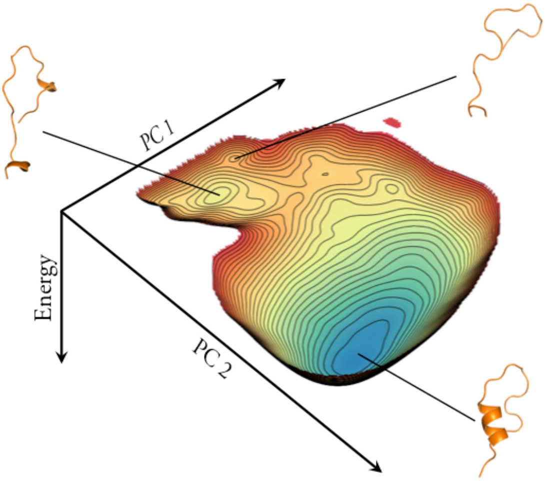


Figure 7.1: Schematic representation of the 3-dimensional free energy landscape of a protein, obtained by using the first two principal components (PCs). Different energy minima contain different protein conformational classes, whose structural similarity is inherently related to how close they are located in the energy landscape.

A SAXS experiment measures the scattering intensity $I(q)$ - upon variation of the scattering angle 2θ - as a function of the scattering vector q[41] defined by:

$$q = \frac{4\pi \sin\theta}{\lambda} \,, \tag{7.9}$$

where λ is the wavelength of the radiation. The corresponding real space distance d is, according to Bragg's law, obtained by the following expression:

$$d = \frac{2\pi}{q} \,. \tag{7.10}$$

The scattering curve $I(q)$ is expressed as:

$$I(q) = F(q) \cdot S(q) \,, \tag{7.11}$$

where $F(q)$ is the **form factor** of the scattering object, containing all information about the shape of the protein; and $S(q)$ is the **structure factor** which contains information on how particles interact with one another. For an ideal solution, that is, a solution diluted to the point where no intermolecular interactions effectively exist, there is no significant contribution from the structure factor ($S(q) = 1$) and the following approximation holds true:

$$I(q) \approx F(q) \,. \tag{7.12}$$

---

[41] The use of q to define something else other than the (partial) charge of an atom or particle is a particular exception of this section about SAXS.

For non-ideal solutions, the structure factor only tends to 1 at medium to high q values. Measurements at different protein concentrations and extrapolation to zero concentration are, therefore, often required to eliminate the contribution of the structure factor on the measured scattering intensity at low angles. The size of the objects that can be analyzed with SAXS typically range from a few to several hundred ångström, and the maximum size of the object is limited by the smallest angle that the instrument can attain for measuring the scattering intensity (Pusey 2002; Receveur-Bréchot & Durand 2012; Svergun & Koch 2003).

In a typical SAXS study, the **scattering intensity curve** (or form factor, recall Eq. 7.12) - where $I(q)$ *vs.* q is followed - is the most common representation of the data, as it represents not only the direct output of the instrument, but also allows the calculation of the size of a particle in the form of the radius of gyration[42]. The interatomic distances r within a protein can be accessed through the **pair-distance distribution function** $P(r)$, which is itself obtained by simply taking the Fourier transform of the scattering curve. This particular representation is rather useful, as it provides a more human-readable medium for the valuable information contained in the scattering curve. A simple visual inspection of $P(r)$ usually provides great insight about the shape, anisotropy and degree of compactness of a protein. Additionally, it also provides another way of determining the radius of gyration, which is often considered superior when studying proteins with extended conformations, as is the case with IDPs (Receveur-Bréchot & Durand 2012). A third and final representation, where $(q \cdot R_g)^2 \cdot I(q)$ is plotted as a function of $q \cdot R_g$, provides an extremely useful way to quickly evaluate the globular nature of a polypeptide chain without requiring the use of theoretical models. In this representation, normally referred to as **Kratky plot**, globular proteins present a maximum value at $q \cdot Rg = \sqrt{3}$ (see blue curve in Figure 7.2), regardless of the size of the protein. Conversely, for a random chain, the curve keeps on rising, eventually reaching a nearly flat region between $q \cdot Rg = 1.5$ and 2, which may eventually keep increasing for more rod-like, rigid polypeptide chains (see red and green curves in Figure 7.2).

The calculation of SAXS data from computer simulations is of great relevance for the validation of protein models and force fields. This is even more important for the specific case of unfolded and disordered proteins, for which X-ray diffraction and NMR techniques are of very limited application, and the information obtained through SAXS is one of the few available bridges between theory and experiment. There are several methods for evaluating the solution scattering of biological molecules from their simulation coordinates, and the main difference among them is on how the solvent is modeled. Older protocols such as CRYSOL (Svergun et al. 1995) - which was first

---

[42] There are different methods for deriving this property from the scattering curve, each having their own *pros* and *cons*, but these will not be discussed here due to their irrelevance for the particular scope of this work. The reader is instead referred to the work of Receveur-Bréchot & Durand (2012) for a succinct presentation of these topics.

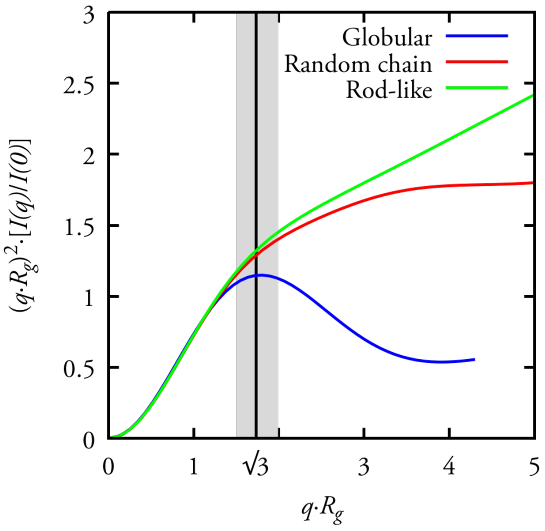


Figure 7.2: Example of the different Kratky plots obtained for proteins belonging to various conformational classes. The vertical black line marks all points with abscissa $\sqrt{3}$. The gray shaded rectangle highlights the area where $1.5 \leq q < 2$.

distributed in 1995[43] - generally treat the solvent as a continuous electron density, where the solvation layer is described as a homogeneous excess electron density, typically 10 % to 15 % higher than that of bulk water (Svergun et al. 1995). According to Chen & Hub (2014), one of the main drawbacks of implicit solvent methods is that these often require two to three free parameters, which are not easily measurable and differ between solutes. In fact, as shown in the Supplementary Information of Paper IV, the default value of the contrast of hydration shell ρ in CRYSOL appears to be ill-suited for use with disordered proteins, and small changes in this value lead to markedly different Kratky plots, which could definitely cause erroneous interpretations. Explicit solvent methods, such as the recently developed WAXSiS (Chen & Hub 2014; Knight & Hub 2015), provide a more accurate model of solvation at the expense of a much higher computational cost. Such methods should, however, be free parameter-free and are required for the calculation of scattering patterns at wider angles (as in wide-angle X-ray scattering, often abbreviated to WAXS), which expands the range of solution X-ray scattering profiles from $4\,\mathrm{nm}^{-1}$ to as much as $50\,\mathrm{nm}^{-1}$, effectively paving the way to a wealth of new structural information (Chen & Hub 2014).

---

[43] A time where the transistor count in a typical microprocessor was between 1 and 5 million, in contrast to the several thousand million transistors of nowadays (Wikipedia 2016b). It is thus understandable, that researchers would generally attempt to avoid considering the many thousand water molecules usually present in a medium sized simulation box when using an explicit water model.

8 Summary of results and outlook

A short summary of the main results of each paper included in this work is presented below. Notice that their respective numbering does not follow a chronological order. These papers are organized according to their particular context instead, and we can envision their division as follows:

- Introduction to the thematic (**Paper I**).
- Development and application of a coarse-grained model for the Monte Carlo simulation of flexible (phospho)proteins, in bulk and near uniformly charged surfaces (**Papers II and III**).
- Presentation of the inherent limitations of standard molecular dynamics force fields for the simulation of IDPs. Assessment of the representativeness of protein disorder models. Application of new "IDP-tailored" force fields and water models to the simulation of disordered proteins of relevance (**Papers IV, V and VI**).

8.1 Paper I

In this first publication, the adsorption mechanism of the histidine rich, unstructured protein Histatin 5 was studied as a function of pH, salt, and multivalent ions. This study combined atomistic molecular dynamics and coarse-grained Metropolis Monte Carlo simulations, as well as classical polymer density functional theory. This multiscale modeling[44] provided a consistent picture in good agreement with experimental data, and the main conclusions were that: (i) proton charge fluctuations promote electrostatic interactions with anionic surfaces through charge regulation; and (ii) specific zinc(II)-histidine binding competes with protons and ensures (an unusually) constant charge distribution over a broad pH interval, which further enhances surface adsorption. When taken together, these points suggest that charge regulation is a significant driving force for the remarkably robust activity of histidine rich antimicrobial peptides.

Despite the success of this study, it became apparent that the calculated protein size, more specifically, the radius of gyration, differed significantly between coarse-grained and atomistic approaches. It was then hypothesized that the "true" R_g should probably lay in-between both models' predictions, which was later confirmed though SAXS measurements (Cragnell et al. 2016). The discrepancy between experiment and the predictions from atomistic simulations would later become an important catalyst for Paper IV.

---

[44] Applying models with different levels of detail (and theory) for the calculation of particular system properties.

8.2 Paper II

Experimental evidence from a previous study (Svensson et al. 2014) showed that the strongly anionic ($\langle Z \rangle \approx -18e$ at pH = 8.5) and flexible phosphoprotein β-casein adsorbs onto negatively charged and hydrophilic silica surfaces. At first, this result appears to be rather counterintuitive due to the fact that there should be a strong electrostatic repulsion between the protein and the surface, as is supported by the application of a coarse-grained model similar to the one used in Paper I, where the only protein-surface interactions considered are of electrostatic nature, modeled according to the Gouy-Chapman theory (recall Sections 2.3.1 and 3.1). However, protein adsorption is a complex process that is controlled by several different mechanisms, e.g., (i) electrostatic interactions between the protein and the surface, and (ii) between adsorbed proteins; (iii) dispersion interactions; (iv) hydration effects; and (v) structural rearrangements of the protein in order to balance conformational chain entropy with energetics. As seen earlier in Sections 2.3.6 and 2.3.7, dispersion interactions are strictly attractive and ubiquitous, and, while individually weak, if present in large numbers may present a significant contribution the the overall system Hamiltonian. Given that β-casein is a long[45] and flexible protein, it was hypothesized that not considering this type of interaction could be the reason behind the discrepancy between experiment and simulation. Thus, a shifted Lennard-Jones potential (recall Section 3.1 and Figure 3.3) was included in the model and it was found that a minimum interaction strength of 2.25 kT was needed in order to promote adsorption and mimic experimental results[46]. Additionally, it was also found that considerable protein net charge fluctuations, due to phosphorylated serine saturation, only have a negligible contribution to the free energy of adsorption.

8.3 Paper III

The coarse-grained model developed in Paper II was used to compare the properties and behavior of β-casein and PRP-1, both in bulk and near a negatively charged surface, in order to evaluate the possibility of using β-casein as a replacement for PRP-1 in pharmaceutical saliva substitutes and, possibly, dental products. Special attention was dedicated to the study of the effect of varying pH, monovalent salt concentration and charge saturation/insaturation of the phosphorylated serine residues (mimicking the binding/release of calcium ions). Both disordered proteins were found to possess very similar electrostatic properties in bulk, specially at physiological pH values and when simulating calcium saturation. Furthermore, when studying surface adsorption it was observed that both proteins attach to the surface in similar manners, relatively to their respective sizes. In particular, the adsorption of both proteins onto the negatively charged surface is strikingly similar under physiological pH and near physiological salt concentrations. On average, however, PRP-1 appears to adsorb more strongly to the surface, while β-casein is generally able to come into closer contact with it. The effect of calcium saturation on surface adsorption was found to be almost negligible for both proteins at high salt concentration.

---

[45] Its sequence contains 209 amino acid residues (recall Table 1).

[46] A following study by Hyltegren et al. (2016), using the model developed here, found that an interaction strength of 2.9 kT is required when simulating Histatin 5, in order to obtain a simulation surface coverage close to experimental reference.

8.4 Paper IV

This study was performed in light on the increasing number of publications - using atomistic molecular dynamics simulations of unfolded and intrinsically disordered proteins - suggesting that standard[47] force fields produce IDP conformational ensembles that are overly collapsed, when compared to experimental reference. Thus, the main goal of this study was to assess the applicability of several (then) state-of-the-art protein force fields for the simulation of Histatin 5, one of the recurrent IDP models used by the Skepö group[48]. The quality of the simulations was assessed in three complementary analyses: (i) protein shape and size comparison with experimental SAXS data obtained by Cragnell et al. (2016); (ii) secondary structure prediction; (iii) (free) energy landscape exploration and conformational class analysis (as discussed in Section 7.3). The results showed that, indeed, standard force fields sample IDP conformations which are too compact, being systematically unable to reproduce experimental evidence such as the form factor, the Kratky plot, and the pair-distance distribution function (recall Section 7.4). Moreover, the consistency of this deviation seems to suggest that the problem may not be mainly due to protein-protein or water-water interactions, whose parameterization varies the most between force fields and water models. In fact, as originally proposed by Best et al. (2014), balanced protein-water interactions appear to be the key to solving this issue, and additional simulations using a modified version of the original TIP4P/2005 water model containing increased protein-water dispersion interactions (recall the TIP4P/2005s water model in Table 2) produces results in very good agreement with experiment.

8.5 Paper V

During the publication process of Paper IV, Piana et al. (2015) introduced a new, four-point water model, called TIP4P-D[49], in which the water dispersion interaction is augmented[50], and the remaining non-bonded parameters are optimized with respect to experimental liquid water properties. Even though there was no opportunity to include this new model in that specific study, it was hypothesized that - judging from the results shown in each respective publication - there was good reason to expect both approaches to produce similar results for the system at hand. To assess the accuracy of this statement, additional molecular dynamics simulations of Histatin 5 were performed using the new water model in conjugation with a standard force field, shown previously to produce poor results when used with the popular TIP3P water model[51]. As predicted, the new simulation results were in excellent agreement with the approach of Best et al. (2014), as studied in Paper IV.

---

[47] Here, "standard" stands for the commonly used protein force fields, not (specifically) developed with IDPs in mind.

[48] Due to its short size, i.e., 24 amino acid residues, and interesting pharmaceutical properties (recall Section 1).

[49] The "D" in the suffix stands dispersion.

[50] Regardless of the specific intervening species, i.e., whether we consider water-water, protein-water or protein-protein dispersion interactions.

Another focal point of this work was to collect and analyze the IDPs that had been studied that far using these new approaches, and investigate how representative of protein disorder they are. With this in mind, an exhaustive analysis of several protein properties of interest, such as the sequence length, amino acid residue content, fraction of charged residues, and net charge per residue, was performed. The results were then compared with different databases for intrinsically disordered and structured proteins. In general terms, the IDPs used to test and validate the new IDP-tailored/aware approaches (Best et al. 2014; Piana et al. 2015) seem to be representative of disordered protein sequences. However, most model proteins appear to be too short in comparison to the average IDP, and their sequences contain a bias toward hydrophilic amino acid residues, with several key order- and disorder-promoting residues being clearly misrepresented. It seems appropriate for future studies to address these issues.

8.6 Paper VI

Histatin 5 is an antimicrobial and disordered protein that acts as the first line of defense against oral candidiasis. It has been shown that conjugation of its active fragment (amino acid residues 4 to 15) with the polyamine spermidine has an even greater candidacidal effect (Tati et al. 2014). However, prior to this study, little to no knowledge about the structure of these conjugates existed. As such, the aim of this study was to characterize the structural properties of Histatin 5_{4-15}-spermidine conjugates by making use of both theoretical and experimental methodologies, as is customary within the Skepö research group. On the theoretical side, apart from the charge parameterization of the conjugate and its implementation in the AMBER ff99SB-ILDN force field, this study is a direct application of the methodology presented in Paper V. On the experimental side, SAXS and circular dichroism measurements were performed. Once more, very good agreement between simulation and experiment was acchieved, suggesting that the force field, water model and parameterization are rather reliable, even for short and strong polyelectrolytes, as is the case with the conjugates. The Histatin 5_{4-15}-spermidine conjugates were found to adopt extended and somewhat rigid conformations in aqueous solution, in contrast to what had been previously hypothesized by Tati et al. (2014), i.e., that the conjugates should adopt relatively compact or globular structures when compared to Histatin 5. No secondary structure was predicted in both aqueous and organic solutions, and the results suggest that the increased antifungal activity of the C-terminal conjugate(s), in comparison to the N-terminal counterpart(s), could be explained by its slightly more extended and rigid conformational ensemble, which allows spermidine to be more exposed to the solvent, thus making it easily accessible for recognition by the polyamine transporters in the cell.

---

[51] For a comparison between water models, please refer back to Section 3.3.

This study also shows how valueable simulations are, given that, for example, the SAXS measurements do not provide information for the entire conjugate, due to the small contrast between the bonded spermidine molecule and the background, i.e., the solvent. Thus, it was from the simulations alone, that it was observed that conjugation with spermidine does not seem to affect the conformational ensemble of the active fragment, as similar regions of coordinate space are sampled when simulating the fragment alone and conjugated to spermidine in any of the four possible variants.

8.7 Outlook

The publications included in this work contain original research performed over the past five years. During this period, and in global terms, considerable progress was made on the modeling and simulation of intrinsically disordered proteins. This is specially true for the particular case of atomistic models and force fields, which despite having enjoyed considerable success with the simulation of folded proteins for quite some time, were recently shown to be inappropriate for the simulation of unfolded and disordered proteins (Lindorff-Larsen et al. 2012; Best et al. 2014; Palazzesi et al. 2014; Piana et al. 2015; Henriques et al. 2015; Rauscher et al. 2015; Ye et al. 2015; Mercadante et al. 2015). While there had been older reports where the atomistic simulation of IDPs was shown to produce overly collapsed conformational ensembles when compared to experimental evidence, it was only until about 2014 that several independent groups started producing more thorough reports, fully dedicated to this subject. It was also around that time that alternative solutions started being proposed, and by mid-August 2015 - the date when the CECAM conference entitled “ Intrinsically Disordered Proteins: Bringing together Physics, Computation and Biology” was held in Zurich, Switzerland - four different research groups had already proposed alternative methods to deal with the clear deficiencies in commonly used simulation models and force fields[52]. Since evidence suggests that the solvent plays a crucial role in shaping the ensembles of intrinsically disordered proteins (Florová et al. 2010), most of these approaches involve modifications to the water parameters, thereby favoring protein–water over protein–protein interactions, which effectively decreases the strength of hydrophobic effect and produces more extended conformations, better matching experimental evidence.

---

[52] (i–ii) The TIP4P/2005s and TIP4P-D water models by Best et al. (2014) and Piana et al. (2015), respectively; (iii) the CHARMM22* force field of Piana et al. (2011) as shown by Rauscher et al. (2015); and (iv) the Kirkwood–Buff derived force field by Mercadante et al. (2015).

In spite of the considerable progress, the literature is still lacking studies where, ideally, all of these new approaches are tested and compared exhaustively using a greater number of unfolded and disordered model proteins, covering all different IDP conformational classes and a spanning a reasonable size range, from oligopeptides to large IDPs containing several hundreds of amino acid residues. Such scientific endeavor would no doubt require significant human and computational resources, but it appears well justified, as greater (and more appropriate) statistical sampling is needed in order to assess how general and robust these approaches are. Additionally, on a more fundamental level, it is equally important to proceed with the academic effort in understanding why folded and disordered proteins behave so differently in aqueous solution, because it is only through a better understanding this topic that we can further improve current models or even develop new ones.

To end, it is important to reiterate the importance of having a reliable model for the simulation of IDPs, as deciphering their molecular mode of action at the structural level remains highly challenging from an experimental point of view (Receveur-Bréchot & Durand 2012). In fact, as shown in Paper VI, the amount of information obtainable from experimental methods is often limited and, sometimes, incomplete. A well proven simulation model could provide a wealth of information at a level of detail which unattainable by any other means, at a fraction of their cost. Take, for example, the study of the temperature-induced collapse of IDPs in aqueous solution. This is a rather interesting and somewhat counter-intuitive phenomenon, whose mechanism is not fully understood. Computer simulations could, in principle, play a major role in aiding its interpretation. However, as reported by Nettels et al. (2009), different models produce different (and often divergent) results when studying the variation of the radius of gyration of IDPs as a function of temperature.

9 References

Abascal, J. L. & Vega, C. 2005, The Journal of Chemical Physics, 123, 234505

Abraham, M. J., Murtola, T., Schulz, R., et al. 2015, SoftwareX, 1, 19

Allen, M. P. & Tildesley, D. J. 1987, Computer Simulation of Liquids, 1st edn. (Oxford University Press)

Bennick, A. 1982, Molecular and Cellular Biochemistry, 45, 83

Berendsen, H. J. 1991, in Computer Simulation in Materials Science, ed. M. Madeleine & P. Vassilis (Springer), 139–155

Berendsen, H. J., Postma, J. P., van, et al. 1984, The Journal of Chemical Physics, 81, 3684

Berendsen, H. J., Postma, J. P., van Gunsteren, W. F., & Hermans, J. 1981, in Intermolecular Forces, ed. P. Bernard (Springer), 331–342

Best, R. B., Zheng, W., & Mittal, J. 2014, Journal of Chemical Theory and Computation, 10, 5113

Bussi, G., Donadio, D., & Parrinello, M. 2007, The Journal of Chemical Physics, 126, 014101

Campos, S. R. & Baptista, A. M. 2009, The Journal of Physical Chemistry B, 113, 15989

Chen, P.-c. & Hub, J. S. 2014, Biophysical Journal, 107, 435

Cragnell, C., Durand, D., Cabane, B., & Skepö, M. 2016, Proteins: Structure, Function, and Bioinformatics, 84, 777

Darden, T., York, D., & Pedersen, L. 1993, The Journal of Chemical Physics, 98, 10089

Das, R. K., Ruff, K. M., & Pappu, R. V. 2015, Current Opinion in Structural Biology, 32, 102

Dunker, A. K., Babu, M. M., Barbar, E., et al. 2013, Intrinsically Disordered Proteins, 1, e24157

Eliezer, D. 2009, Current Opinion in Structural Biology, 19, 23

Evans, D. F. & Wennerström, H. 1999, The Colloidal Domain: Where Physics, Chemistry, Biology, and Technology Meet, 2nd edn. (Willey-VCH)

Evers, C. H., Andersson, T., Lund, M., & Skepö, M. 2012, Langmuir, 28, 11843

Farrell, H., Jimenez-Flores, R., Bleck, G., et al. 2004, Journal of Dairy Science, 87, 1641

Florová, P., Sklenovský, P., Banáš, P., & Otyepka, M. 2010, Journal of Chemical Theory and Computation, 6, 3569

Frenkel, D. & Smit, B. 2002, Understanding Molecular Simulation: From Algorithms to Applications, 2nd edn. (Academic Press)

Ghanty, T. K., Staroverov, V. N., Koren, P. R., & Davidson, E. R. 2000, Journal of the American Chemical Society, 122, 1210

Greberg, H., Kjellander, R., & Åkesson, T. 1996, Molecular Physics, 87, 407

Hay, D., Bennick, A., Schlesinger, D., et al. 1988, Biochemical Journal, 255, 15

Henriques, J., Cragnell, C., & Skepö, M. 2015, Journal of Chemical Theory and Computation, 11, 3420

Henriques, J. & Skepö, M. 2016, Journal of Chemical Theory and Computation, 12, 3407

Hess, B., Kutzner, C., Van Der Spoel, D., & Lindahl, E. 2008, Journal of Chemical Theory and Computation, 4, 435

Hill, T. L. 1986, An Introduction to Statistical Thermodynamics (Dover Publications, Inc., New York)

Hockney, R. W. & Eastwood, J. W. 1988, Computer Simulation Using Particles (CRC Press)

Hyltegren, K., Nylander, T., Lund, M., & Skepö, M. 2016, Journal of Colloid and Interface Science, 467, 280

Isaacs, E., Shukla, A., Platzman, P., et al. 1999, Physical Review Letters, 82, 600

Israelachvili, J. N. 2011, Intermolecular and Surface Forces, 3rd edn. (Academic Press)

Jackson, M. B. 2006, Molecular and Cellular Biophysics (Cambridge University Press)

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. 1983, The Journal of Chemical Physics, 79, 926

Jorgensen, W. L. & Madura, J. D. 1985, Molecular Physics, 56, 1381

Kjellander, R. 2012, Statistical thermodynamics course Compendium (University of Gothenburg)

Knight, C. J. & Hub, J. S. 2015, Nucleic Acids Research, 43, W225

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., & Shaw, D. E. 2012, Journal of the American Chemical Society, 134, 3787

London, F. 1937, Trans. Faraday Soc., 33, 8b

Lund, M. & Jönsson, B. 2013, Quarterly Reviews of Biophysics, 46, 265

Martín, M. B. 2012, Boundary conditions, `dingercatadventures.blogspot.se`

McLachlan, A. 1965, Discussions of the Faraday Society, 40, 239

Mercadante, D., Milles, S., Fuertes, G., et al. 2015, The Journal of Physical Chemistry B, 119, 7975

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, The Journal of Chemical Physics, 21, 1087

Napeñas, J. J., Brennan, M. T., & Fox, P. C. 2009, Odontology, 97, 76

Nettels, D., Müller-Späth, S., Küster, F., et al. 2009, Proceedings of the National Academy of Sciences, 106, 20740

Nosé, S. & Klein, M. 1983, Molecular Physics, 50, 1055

Oppenheim, F., Xu, T., McMillian, F., et al. 1988, Journal of Biological Chemistry, 263, 7472

Palazzesi, F., Prakash, M. K., Bonomi, M., & Barducci, A. 2014, Journal of Chemical Theory and Computation, 11, 2

Parrinello, M. & Rahman, A. 1981, Journal of Applied physics, 52, 7182

Piana, S., Donchev, A. G., Robustelli, P., & Shaw, D. E. 2015, The Journal of Physical Chemistry B, 119, 5113

Piana, S., Lindorff-Larsen, K., & Shaw, D. E. 2011, Biophysical Journal, 100, L47

Pronk, S., Páll, S., Schulz, R., et al. 2013, Bioinformatics, 29, 845

Pusey, P. N. 2002, in Neutron, X-rays and Light: Scattering Methods Applied to Soft Condensed Matter, 1st edn., ed. T. Zemb & T. Lindner (North-Holland), 3–22

Rauscher, S., Gapsys, V., Gajda, M. J., et al. 2015, Journal of Chemical Theory and Computation, 11, 5513

Rauscher, S. & Pomès, R. 2010, Biochemistry and Cell Biology, 88, 269

Receveur-Bréchot, V. & Durand, D. 2012, Current Protein and Peptide Science, 13, 55

Ribadeau, D. B., Brignon, G., Grosclaude, F., & Mercier, J. 1972, European Journal of Biochemistry/FEBS, 25, 505

Schenkels, L. C., Veerman, E. C., & Amerongen, A. V. N. 1995, Critical Reviews in Oral Biology & Medicine, 6, 161

Svensson, O., Kurut, A., & Skepö, M. 2014, Food Hydrocolloids, 36, 332

Svergun, D. I., Barberato, C., & Koch, M. H. J. 1995, Journal of Applied Crystallography, 28, 768

Svergun, D. I. & Koch, M. H. J. 2003, Reports on Progress in Physics, 66, 1735

Tati, S., Li, R., Puri, S., et al. 2014, Antimicrobial Agents and Chemotherapy, 58, 756

Tironi, I. G., Sperb, R., Smith, P. E., & van Gunsteren, W. F. 1995, The Journal of Chemical Physics, 102, 5451

Tompa, P. 2002, Trends in Biochemical Sciences, 27, 527

Tompa, P. 2011, Current Opinion in Structural Biology, 21, 419

Tompa, P. 2012, Trends in Biochemical Sciences, 37, 509

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. 2008, Annual Review of Biophysics, 37, 215

van der Spoel, D., Lindahl, E., Hess, B., & the GROMACS development team. 2014, GROMACS User Manual version 4.6.7, www.gromacs.org

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. 2004, Journal of Molecular Biology, 337, 635

Wei, G.-X. & Bobek, L. A. 2005, Antimicrobial Agents and Chemotherapy, 49, 2336

Weinhold, F. 1997, Journal of Molecular Structure: THEOCHEM, 398, 181

Wikipedia. 2016a, Cube — Wikipedia, The Free Encyclopedia

Wikipedia. 2016b, Moore's law — Wikipedia, The Free Encyclopedia

Wikipedia. 2016c, Rhombic dodecahedron — Wikipedia, The Free Encyclopedia

Williams, R., Obradović, Z., Mathura, V., et al. 2001, in Pacific Symposium on Biocomputing, Vol. 6, 89–100

Wong, R. & Bennick, A. 1980, Journal of Biological Chemistry, 255, 5943

Wright, P. E. & Dyson, H. J. 1999, Journal of Molecular Biology, 293, 321

Xu, T., Levitz, S., Diamond, R., & Oppenheim, F. 1991, Infection and Immunity, 59, 2549

Ye, W., Ji, D., Wang, W., Luo, R., & Chen, H.-F. 2015, Journal of Chemical Information and Modeling, 55, 1021

Scientific publications

Author contributions

Paper i: Role of histidine for charge regulation of unstructured peptides at interfaces and in bulk

Involved in planning and writing the manuscript. Performed the atomistic simulations and respective analyses.

Paper ii: A coarse-grained model for flexible (phospho)proteins: adsorption and bulk properties

Involved in planning and writing the manuscript. Developed and implemented the simulation algorithm and analysis routines with input from the co-author. Performed the simulations and analyses. Responsible for the submission and revision process.

Paper iii: *In silico* physicochemical characterization and comparison of two intrinsically disordered phosphoproteins: β-casein and acidic PRP-1

Initiated, planned and formulated the scientific question and hypothesis of the project together with the co-authors. Assisted the second author in performing the bulk of the simulations and analyses, using own algorithms and routines. Performed a small part of the simulations and analyses. Wrote the manuscript together with the co-authors. Responsible for the submission and revision process.

Paper iv: Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment

Initiated, planned and formulated the scientific question and hypothesis of the project together with the co-authors. Performed the simulations. Defined, implemented and performed the simulation analyses. Wrote the manuscript with input from the co-authors. Responsible for the submission and revision process.

Paper v: Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models

Initiated, planned and formulated the scientific question and hypothesis of the project together with the co-author. Performed the simulations and analyses. Wrote the manuscript with input from the co-author. Responsible for the submission and revision process.

Paper vi: Structural characterization of Histatin 5-spermidine conjugates: a combined experimental and theoretical study

Initiated, planned and formulated the scientific question and hypothesis of the project together with the co-authors. Performed the force field parameterization and assisted the first author in performing and analyzing the simulations. Wrote the manuscript together with the co-authors. Responsible for the submission process.

Printed by Media-Tryck, Lund University

This work is primarily about the development, validation and application of computer simulation models for intrinsically disordered proteins, both in solution and in the presence of uniformly charged, ideal surfaces. The models in question are either coarse-grained or atomistic in nature, and their applications are dependent on the specific purpose of each study. Both, Metropolis Monte Carlo and molecular dynamics simulations were used to run these models.

To me, however, this work is about something completely different. It is the culmination of a long and arduous five year period of hectic activity, learning and maturation. In total, I performed approximately 140 CPU years worth of simulations, completed 11 courses, participated in 8 international conferences and workshops - with 6 poster presentations, 2 talks and 2 awards - and travelled to 6 different countries. On top of that, I was responsible for the exercise sessions of the KEMB08 course for two years in a row, co-supervised a student twice, published 5 articles, submitted another and am still working on 2 other projects, which are to be concluded and published in the near future. Last, but definitely not least, I became a father midway through all of this. All in all, I would say it was quite an enriching experience and I now hope that the knowledge gathered throughout my doctoral studies, as presented here, can be of use to others.

ISBN 978-91-7422-489-4 (print)
ISBN 978-91-7422-490-0 (pdf)

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University