# On the Thermodynamic Solvation of Biomolecules in Solution

Hervö Hansen, Stefan

2021

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

# On the Thermodynamic Solvation of Biomolecules in Solution

STEFAN HERVØ-HANSEN | DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

ISBN: 978-91-7422-832-8

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University

9 789174 228328

# On the Thermodynamic Solvation of Biomolecules in Solution

# On the Thermodynamic Solvation of Biomolecules in Solution

by Stefan Hervø-Hansen


LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisors: Prof. Dr. Mikael Lund
Faculty opponent: Prof. Dr. David Mobley

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in lecture hall C at Kemicentrum, Department of Chemistry on Friday, the 15th of October 2021 at 13:00.

Organization
**LUND UNIVERSITY**
Department of Chemistry
Box 124
SE–221 00 LUND
Sweden

Document name
**DOCTORAL DISSERTATION**
Date of disputation
2021-10-15
Sponsoring organization

Author(s)
Stefan Hervø-Hansen

Title and subtitle
On the Thermodynamic Solvation of Biomolecules in Solution:

Abstract
The topic solvation thermodynamics is an important aspect of chemistry, dealing with the effects introduced by solvents onto solutes. In particular, biological systems are highly heterogeneous in their choice of solvent typically characterized by either being in a polar or non-polar environment. For example, the cytoplasm of cells constituting the internal environment of cells is an aqueous solvent, whereas the membrane, being the boundary separating the cells from its surroundings, is an example of a lipid solvent. In addition to the main solvent, the majority of biological solvents also contain co-solvents such as ions, including ATP which can be found to be on the 10 mM cellular concentration scale, or monovalent ions such as potassium, sodium, chloride, and not to forget free amino acids. While the previously mentioned examples are important in their own regards for oxidative phosphorylation, nerve cell communication, and construction of proteins respectively, to mention a few examples, their role as co-solvents can also greatly affect the stability and solubility of molecular matter.

In this work, we will investigate the properties underlying the solvation of molecular matter utilizing statistical thermodynamics and molecular simulations. In specific by using molecular simulations we can determine atomistic properties for systems of interest, and via statistical thermodynamics relate these properties to experimental observables. These observables may either be mechanical properties addressing the behavior of molecular matter at a given state, or they may be state functions that describe the changes in energetics and entropy for the molecular matter changing. Within solvation thermodynamics, one of the most important state functions is the chemical potential and highly related solvation free energy describing the free energy of adding a solute particle to the system and thus quantifies the reversible work between the solute and solvent upon introducing the particle, and thus multiple methods are discussed how to obtain this quantity.

The findings of the presented research include among others reflections upon the solubility of salt bridges in proteins, and counter-intuitive cation-cation enthalpic attraction due to changes in ion solvations induced via a host molecule. Furthermore, the research also addresses the regulation of co-solvent-induced aggregation. Last, but not least, the total thermodynamic decomposition of caffeine solvation in electrolyte solutions, utilizing energy-representation theory of solvation to unveil the mechanism of anion-specific processes, and demonstrating the capabilities of the method to unlock solvation properties to optimize and rationally design future systems.

Key words
Solubility, Solvation, Aggregation, Statistical thermodynamics, Molecular dynamics, Monte Carlo simulations, Free energy calculations, Energy-representation theory of solvation, Proteins, Salt bridges, Caffeine.

Classification system and/or index terms (if any)

Supplementary bibliographical information
Language
English

ISSN and key title
ISBN
978-91-7422-832-8 (print)
978-91-7422-833-5 (pdf)

Recipient's notes
Number of pages
314
Price
Security classification

# On the Thermodynamic Solvation of Biomolecules in Solution

by Stefan Hervø-Hansen


LUND
UNIVERSITY

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

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

**Cover illustration front:** Illustration of sodium being hydrated by water accompanied with the release of heat as it is one of the characteristic steps in the calculation of the enthalpy change of solvation.

**Cover illustration back:** A cup of coffee accompanied by a selection of anions. Related to the research conducted in Paper V and VI.

**Funding information:** The thesis work was financially supported by Swedish Research Council; the Swedish Foundation for Strategic Research; the European Research Council through the PIPPI consortium; and the Royal Physiographic Society of Lund.



Faculty of Science, Department of Chemistry

ISBN: 978-91-7422-832-8 (print)
ISBN: 978-91-7422-833-5 (pdf)

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

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

Printed matter
3041 0903

MADE IN SWEDEN

*Dedicated to my parents*
*Linda and Karsten Hervø Hansen*

*And to the memory of*
*John Wells Hervø (1936-2020)*

# Contents

## List of publications

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

I **Charge Interactions in a Highly Charge-depleted Protein**

**S. Hervø-Hansen**, C. Højgaard, K. E. Johansson, Y. Wang, J. Wahni, D. Young, J. Messens, K. Teilum, K. Lindroff-Larsen, J. R. Winther
Journal of the American Chemical Society, 2021, 143, 6, pp. 2500–2508

II **Systematic Electrostatic Perturbation of a Charge-depleted Protein: Correlation between Protein Solubility and Electrostatics**
**S. Hervø-Hansen**, C. Højgaard, J. R. Winther, M. Lund, N. Matubayasi
Manuscript

III **Counter Intuitive Electrostatics upon Metal Ion Coordination to a Receptor with Two Homotopic Binding Site**
V. Aspelin, C. S. Arribas, **S. Hervø-Hansen**, Björn Stenqvist, R. Chudoba, K. Wärnmark, M. Lund
Submitted

IV **Total Description of Intrinsic Amphiphile Aggregation: Calorimetry Study and Molecular Probing**
R. Fernandez-Alvarez, Ž. Medoš, Z. Tošner, A. Zhigunov, M. Uchman, **S. Hervø-Hansen**, M. Lund and M. Bešter-Rogač, P. Matějíček
Langmuir, 2018, 34, 47, pp. 14448–14457

V **Anion-Cation Contrast of Caffeine Solvation in Salt Solutions**

**S. Hervø-Hansen**, M. Lund, N. Matubayasi
Submitted

VI **A Surface Area Description of Salting-in and Salting-out of Caffeine**

**S. Hervø-Hansen**, J. Polák, M. Tomandlová, J. Dzubiella, J. Heyda, M. Lund
Manuscript

VII **Impact of Arginine–Phosphate Interactions on the Reentrant Condensation of Disordered Proteins**
S. Lenton, **S. Hervø-Hansen**, M. D. Tully, M. Lund, M. Skepö
Biomacromolecules, 2021, 22, 4, pp. 1532–1544

# Preface

A popular discussion question upon starting at the university within the biological sciences is how one would define "life". This is a difficult question to answer and usually resolves in a good debate. One reason why life is so hard to define, is because it is not a substance, but a process. Consequently, it is not uncommon to define life based on the characteristics of life. Examples of such characteristics include the possibility of reproduction, the ability to maintain homeostasis, to possess metabolism, and many more. Another characteristic of life that is a little less obvious, but related to homeostasis, is the ability to separate the system from the surroundings. For example bacteria and human cells separate them self using a cell membrane constructed from lipids, while viruses separate them self using a protein capsid. If these barriers are broken, the organisms will die. This is where the concept of solubility and desolvation becomes essential.

In these dire times with COVID-19 causing havoc around the world, it has once again become crucial to recognize the importance of good hygiene and methods how to achieve it. One major effort is the usage of hand sanitizer in addition to the washing of hands with soap, with the scientific community aiding in achieving such by public outreach, production of chemicals, and research for new knowledge. For example, it can be mentioned the large amounts of hand sanitizer produced at Kemicentrum in Lund, Sweden to aid the places where it was needed the most.

For a long time, the main aim of this work has been unclear, however, given my stay in Japan and the following COVID-19 crisis, I realized I wanted to illustrate the importance of solvation thermodynamics, using model systems involved in the many processes of life. The main key questions are in particular: 1. How and why cosolvent affects the physical and chemical equilibrium involving molecular matter. 2. How can we quantify and qualify the effects total and individual contributions of cosolvent on molecular equilibria. 3. How and to which degree can we alter molecular matter to obtain desired solvation properties. To address the first question we investigated various systems at different scales of size ranging from small molecules like caffeine and cobaltabisdicarbollide to proteins like lysozyme, histatine 5, and a cellulose-binding domain from *Cellulomonas fimi*. The second question is mostly related to method development, with the main methods in this work being molecular simulations for the sampling of the configurational space and free energy calculation methods to access the spontaneity of given reactions.

The underlying framework, to connect molecular simulations to experimental observations, is that of statistical thermodynamics. Therefore the success of this work relies on the understanding and simultaneous research and development of statistical thermodynamics and hence takes its fair share of coverage in this work.

Stefan Hervø-Hansen - Reflections on a plane to Finland in 2020.

# Acknowledgements

The journey of my Ph.D. education starts with my coworker during my master's degree, **Anil Kurut**, suggesting I apply for a Ph.D. position at Lund University with her supervisor from her Ph.D. degree. I was asked to come to Sweden for an "informal" interview with the journey being governed by canceled trains and buses causing a lot of problems reaching Lund University, almost as if fate was trying to tell me not to go. Upon arrival, I got to meet **Mikael Lund**, with whom I over lunch discussed programming, simulations, proteins, and being Danish in Sweden. Thanks to your supervision I have been able to grow much more as a scientist and researcher than I could ever have imagined. Simultaneously I also got to meet **Marie Skepö**, who became my co-supervisor. She is a brilliant female scientist who always had my back whenever life felt too hard. Rumors have it, she happily volunteered to be my co-supervisor, however, it was never confirmed. Given her major experience in protein chemistry, I can only imagine it to be true.

With my employment, I was positioned in an office with two talented researchers; **Björn Persson**, who was always there to provide guidance on life and research, and **Coralie Pasquier**, a cat and protein loving woman, who I shared my first Ph.D. project with and later apartment in Åkarp. I can honestly say it was never boring with you two around! Upon starting my courses, I got merged with **Samuel Stenberg** and **Vidar Aspelin** forming a study group of the three of us. Together we explored the amazing world of statistical thermodynamics, simulations, and other fundamental yet advanced topics we found interesting. Thanks to the two of you, the driest and complicated mathematical expressions and abstract theories have turned into fun and excitement, partly with the help of alcohol. I will remember those days as some of the best days in my time as a Ph.D. student.

During my Ph.D. studies, I had a secondment in Osaka, Japan under the supervision of **Nobuyuki Matubayasi**. My time in Japan was crucial for my development and finally made me settled my mind for the topic of my Ph.D. studies to be solvation thermodynamics. I thank you all deeply for the kindness and hospitality you and your group have shown me. Ough, I mean 大へんたいへんお世話せわになり、ありがとうございました。

Along with my Ph.D. studies I also engaged in playing badminton to which I need to mention **Björn Stenqvist**... my mortal (I think) nemesis in badminton and an electrostatics genius. Thank you for listening to my findings and always having time for me and obviously for always reminding me, that I can improve at badminton after you destroyed me.

Finally, I like to thank **everyone in the Division of Theoretical Chemistry**. Every one of you has been contributing to shaping my life during my Ph.D. and I am grateful for all of it. Lastly, I wish to thank **my parents: Linda and Karsen Hervø Hansen** for their endless support from the other side of Øresund.

## Popular summary in English

It is generally known that water and oil cannot be mixed favorably, but that one would instead create a phase separation characterized by having an upper oil phase and a lower aqueous phase. As first stated, this observation is well known, while the mechanism of this phase separation is less well known, unless one is familiar with the interactions that stabilize an oil phase and an aqueous phase and is familiar with the most basic thermodynamics necessary to describe whether processes are spontaneous or not. This is the essence of this work and what is meant by the field of "solvation thermodynamics"; to develop theories and methods to characterize the forces driving system and molecules of interest toward a characteristic state due to the surrounding solvent.

In addition to the problem of whether liquids can spontaneously mix or phase separate, the thermodynamics of solvation also addresses issues such as protein folding, how stereoselective catalysis for drug synthesis can be achieved, the optimization of the performance of an electrochemical capacitor, and the aggregation of molecular substances such as proteins important for diseases such as Parkinson's and Alzheimer's disease. In this thesis, we address in particular the questions: (**I**) how the addition of salt alters caffeine's interactions in aqueous solutions. (**II**) How the addition of phosphate-containing chemicals leads to the aggregation of proteins. (**III**) How the relationship between electrostatic interactions between ions and their corresponding solubility can be described from a physical and thermodynamic perspective.

The method chosen to address the above issues is via "computer experiments", which is in contrast to the traditional perception of chemists performing experiments in white lab coats in a laboratory. Using computer simulations, we can follow the positions and velocities of the individual atoms to create insight into how matter behaves at the atomic level due to solvation effects, which is otherwise almost unattainable by traditional experimental methods. Despite the great potentials given our choice of method, there are equally great challenges. One of the major challenges is to mimic systems studied in the traditional laboratory, which include calibrating molecules' external interactions with other surrounding molecules and adjusting the geometry of molecules by calibrating molecules' internal interactions. Another significant challenge is the sorting and use of the enormous amounts of data that are created in a simulation, that must be used to find the driving forces responsible for inducing changes in chemical systems.

Using simulations and statistical (solvation) thermodynamics, we show how to consistently isolate the effect of the individual solvent molecules and their influence on molecules of interest. Furthermore, we demonstrate how simulations and statistical thermodynamics can be used to interpret experimental data and thus be included as an essential tool for understanding how our world works and operates.

## Populærvidenskablig sammenfatning på dansk

Det er almen kendt, at vand og olie ikke favorabelt kan mikses. I stedet vil man skabe en faseseperation kendetegnet ved at have en øvre olie fase og en nedre vandig fase. Som først sagt er denne observation almindeligt kendt, mens mekanismen for denne fase-separation er mindre almen velkendt med mindre man er bekendt med de interaktioner, der stabiliserer en olie fase og en vandig fase og er bekendt med den mest basale termodynamik, hvilket er nødvendigt til at beskrive om processer er spontane eller ej. Dette er essessen i dette arbejde, og hvad menes med feltet "opløsligheds termodynamik"; at udvikle teorier og metoder til at karakterisere de kræfter, der driver systemer og molekyler af interesse mod en karakteristisk tilstand på grund af det omgivende solvent.

Foruden problematikken hvorvidt væsker spontant kan mikses eller faseseparere, adresserer feltet opløsligheds termodynamik også problemstillinger såsom protein foldning, hvorledes stereoselektiv katalyse til lægemiddelsyntese kan opnås, optimeringen af ydeevnen for en elektrokemisk kondensator og aggregeringen af molekylære stoffer, såsom proteiner der kan lede til sygedomme som Parkinsons- og Alzheimers sygdom. I denne tese adresserer vi særligt spørgsmålene: (**I**) Hvordan additionen af salt ændrer koffeins vekselvirkninger i vandige opløsninger. (**II**) Hvordan additionen af fosfatholdige kemikalier leder til aggregeringen af proteiner. (**III**) Hvordan forholdet mellem elektrostatiske vekselvirkninger mellem ioner og deres korresponderende opløslighed kan beskrives fra et fysisk og termodynamisk perspektiv.

Den valgte metode til at adressere de ovenstående problemstillinger er via "computereksperimenter", hvilket er i kontrast til den traditionelle forestilling om kemikere, der udfører eksperimenter i hvide kitler i et laboratorium. Ved at bruge computersimuleringer kan vi følge de individuelle atomers positioner og hastigheder til at skabe indblik i, hvorledes stof opfører sig på et atomistisk niveau på grund af solvatiseringseffekter, hvilket er ellers næsten uopnåeligt ved traditionelle eksperimentelle metoder. På trods af de store potentialler givet vores metodevalg er der ligeliges store udfordringer. En af de væsentlige udfordringer er at efterligne systemer, der bliver studeret i det traditionelle laboratorium, hvilket blandt andet involverer at kalibrere molekylers eksterne interaktioner med andre omgivende molekyler samt justere molekylers geometri ved at kalibrere molekylers interne interaktioner. En anden væsentlig udfordring er sorteringen og brugen af de enorme mængder af data, der bliver skabt i en simulering, der skal bruges til at finde drivkræfterne bag forandringer i kemiske systemer.

Ved brug af simuleringer og statistisk (opløsligheds) termodynamik viser vi, hvordan man konsekvent kan isolere effekten af de individuelle solventmolekylers indflydelse på drivkræfterne i ændringen af molekyler, som har interesse. Yderemere demonstrerer vi, hvordan simuleringer og statistisk termodynamik kan bruges til at fortolke eksperimentelle data og

dermed indgår som et essentielt værktøj til at forstå hvorledes vores verden virker og opererer.

## Populärvetenskaplig sammanfattning på svenska

Det är allmänt känt att vatten och olja inte kan blandas fördelaktigt utan att man istället skulle skapa en fasseparation som kännetecknas av att ha en övre oljefas och en nedre vattenfas. Som nämnts först är denna observation allmänt känd, medan mekanismen för denna fasseparation är mindre allmänt känd såvida man inte är bekant med interaktionerna som stabiliserar en oljefas och en vattenfas och är bekant med den mest grundläggande termodynamiken som är nödvändig för att beskriva processer är spontana eller inte. Detta är kärnan i detta arbete och vad som menas med fältet "löslighetstermodynamik"; att utveckla teorier och metoder för att karakterisera de krafter som driver systemet och molekyler av intresse mot ett karakteristiskt tillstånd på grund av det omgivande lösningsmedlet.

Förutom problemet med att vätskor spontant kan blandas eller fassepareras, tar lösligheten termodynamik också upp frågor som proteinvikning, hur stereoselektiv katalys för läkemedelssyntes kan uppnås, optimering av prestandan hos en elektrokemisk kondensator och aggregationen molekylära ämnen såsom proteiner, såsom Parkinsons och Alzheimers sjukdom. I denna avhandling behandlar vi särskilt frågorna: (**I**) hur tillsatsen av salt förändrar koffeininteraktioner i vattenlösningar, (**II**) hur tillsatsen av fosfatinnehållande kemikalier leder till aggregering av proteiner och (**III**) hur förhållandet mellan elektrostatiska interaktioner mellan joner och deras motsvarande löslighet kan beskrivas ur ett fysiskt och termodynamiskt perspektiv.

Metoden som valts för att ta itu med ovanstående frågor är via "datorexperiment", vilket står i kontrast till det traditionella begreppet kemister som utför experiment i vita rockar i ett laboratorium. Med hjälp av datasimuleringar kan vi följa positionerna och hastigheterna för de enskilda atomerna för att skapa insikt i hur materia beter sig på atomnivå på grund av solvationseffekter, vilket annars är nästan ouppnåeligt med traditionella experimentella metoder. Trots de stora potentialerna med tanke på våra metodval finns det lika stora utmaningar. En av de största utmaningarna är att efterlikna system som studerats i det traditionella laboratoriet, som inkluderar kalibrering av molekylers externa interaktioner med andra omgivande molekyler och justering av molekylernas geometri genom kalibrering av molekylers interna interaktioner. En annan viktig utmaning är sorteringen och användningen av de enorma mängder data som skapas i en simulering som måste användas för att hitta drivkrafterna bakom förändringar i kemiska system.

Med hjälp av simuleringar och statistisk (löslighet) termodynamik visar vi hur man konsekvent isolerar effekten av de enskilda typer av molekyler som utgör lösningsmedlet har inflytande på intressanta molekyler. Dessutom demonstrerar vi hur simuleringar och statistisk termodynamik kan användas för att tolka experimentdata och därmed inkluderas som ett viktigt verktyg för att förstå hur vår värld fungerar och fungerar.

# Chapter 1

# Theoretical Foundations of Statistical Thermodynamics

*To steal ideas from one person is plagiarism; to steal from many is research.*
— Steven Wright

The field of science is associated with the attempt of achieving certainty for the processes governing our world, allowing us to explain phenomena while simultaneously predicting the past and future outcomes of certain events. Consequently, many formulas and theories have been built and postulated to achieve complete certainty of systems. However, with the discovery of quantum mechanics, it has been discovered, nature cannot be described solely by classical mechanics. Rather nature seems to be governed by uncertainty at a fundamental level leaving us only to predict probability distributions. It is conceivably that the transition from classical mechanics to quantum mechanics may not be considered a paradigm shift as it was described by Thomas Kuhn, due to that the two branches of science can be considered an extension of one another. The transition in terms of our limits to understand and predict events happening in the world we populate certainly is! While quantum mechanics applies to the subatomic world and classical mechanics is commonly applied to the macroscopic world, the two branches of mechanics are interconnected by **statistical mechanics**. As a consequence, in this chapter, we will introduce the fundamental and necessary theories, concepts, and approximations required to gain sufficient knowledge to approach the topic of solvation thermodynamics. In particular we will discuss concepts such as *ensemble theory*, the *ensemble distribution functions* and introduce the concepts of *entropy* and *free energy*. While the experienced reader may be familiar with these concepts, the following text will put emphasis on maintaining a strong connection between mathematics and physical interpretation. One such example is the concept such as entropy, which can be highly

elusive in terms of detailed understanding even by experienced scientists.

## 1.1 Introduction & a Word of Caution

At the time of the development of thermodynamics, the microscopic origin of the macroscopic thermodynamic observables was unknown. Thus causing the thermodynamic laws we all know so well today to be regarded as phenomenological laws. Today we take atoms and atomic theory for granted and can be used to fundamentally explain the microscopic origin of thermodynamics. However, in the early development of statistical thermodynamics by Ludwig Boltzmann and contemporaries, describing a gas enclosed in a volume as a huge collection of ultra-small particles in constant motion and constantly colliding with one another was revolutionary. Due to the huge number of particles a deterministic approach would be impossible (which is still true today), thus causing Boltzmann in a moment of genius to instead realize that one could utilize the probability of individual particles to be traveling at certain speeds and directions to build working theories that match experiments with great accuracy. Unfortunately, due to the governing paradigm at the time being determinism and the concept of atoms was consider fictitious calculation devices, his theory was met with intense hostility causing Boltzmann to take his own life in 1906 just one year after Albert Einstein inevitably proved the existence of atoms. As a consequence and for the safety of the reader, the topic of statistical thermodynamics is best studied open-minded.

## 1.2 Equilibrium Ensembles & Averages

For a system in thermodynamic equilibrium, the thermodynamic state (macrostate) of the system is unchanging over time given the system is not disturbed. If we could imagine a large system of particles (on the molar scale) enclosed in a volume, the system could be further subdivided into smaller volumes (nanomolar scale). Any thermodynamic variable computed for the sub-volumes would yield a distribution of values that together would form a *phase average*. However, given the statement that a system does not change thermodynamic state and that particles at non-zero Kelvin are in constant motion, as emphasized by Boltzmann, it should be sufficient to simply observe the time evolution of a single subvolume into the configurations observed that made up the ensemble average. The ensemble of configurations observed from the time-evolution is called the *time average*. The *ergodic hypothesis* put forth by Boltzmann and Maxwell states: the phase average and time average are equivalent. Mathematically the average quantity $f$ can at equilibrium be calculated as

$$\langle f \rangle = \int f(\vec{p},\vec{q})\rho(\vec{p},\vec{q})\mathrm{d}\vec{p}\mathrm{d}\vec{q} = \lim_{\mathcal{T}\to\infty}\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} f(t)\mathrm{d}t, \tag{1.1}$$

where $\mathcal{T}$ is the duration of the observation and $\rho(p,q)$ is the probability of observing a specific microstate with momentum $p$ and position $q$. This relationship is crucial to provide a physical interpretation of statistical thermodynamics and its link to experiments.[83]

## 1.3 The Ensemble Distribution Function & Liouville's theorem

In Eq. 1.1 we how saw the average could be expressed via a time average and an ensemble average. This relationship is resting on the assumption that the probability did not depending on time, and therefore constant throughout the observation. One prerequisite to obtain that result is Liouville's theorem, which states that for a given macrostate, the phase space probability density is constant in time. Mathematically we can write the statement as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\partial\rho}{\partial t} + \sum_{i=1}^{3N}\left(-\frac{\partial\rho}{\partial p_i}\frac{\partial H}{\partial q_i} + \frac{\partial\rho}{\partial q_i}\frac{\partial H}{\partial p_i}\right) = 0, \tag{1.2}$$

where $H$ is the Hamiltonian, $p_i$ is the momentum, and $q_i$ is the position of particle $i$. Liouville's theorem was recognized by Josiah Gibbs as perhaps the most fundamental relationship of statistical thermodynamics with the theorem being equally applicable in equilibrium and non-equilibrium statistical thermodynamics.[42] If we now choose to *define* an equilibrium system as one in which the density of states is time-independent, meaning that the partial derivative of the phase space density with respect to time is zero, the second term must also yield zero:

$$\sum_{i=1}^{3N}\left(-\frac{\partial\rho}{\partial p_i}\frac{\partial H}{\partial q_i} + \frac{\partial\rho}{\partial q_i}\frac{\partial H}{\partial p_i}\right) = 0. \tag{1.3}$$

An important feature obtainable from Eq. 1.3 is, that the phase space probability density depends exclusively on the Hamiltonian and not time explicitly. The challenge is now to solve Eq. 1.3 for the probability density distribution, $\rho$, under different constraints leading to the probability density distributions for various statistical equilibrium ensembles.

## 1.4 The Microcanonical Ensemble

The simplest solution to Eq. 1.3 is under the constraint of constant energy, volume, and number of particles donated the microcanonical ($NVE$) ensemble. In this ensemble each possible microstate possess equal probability, thus meaning that the probability density distribution function is a constant. The microcanonical density distribution function on the form proposed by Landau & Lifshitz[61] is given by

$$\rho_{NVE}(\vec{p},\vec{q}) = \frac{1}{\omega(E)}\delta\left[E - H(\vec{p},\vec{q})\right], \tag{1.4}$$

where $\delta$ is the Dirac delta function ensuring the microcanonical density distribution function can be written as a continuous function, and $\omega(E)$ is a normalization constant. For probability density functions the normalization constant should have the property of ensuring the area under its graph is equal to one:

$$1 \overset{\text{Def}}{=} \int \rho_{NVE}(\vec{p},\vec{q})\mathrm{d}^N\vec{p}\,\mathrm{d}^N\vec{q} = \frac{1}{\omega(E)}\int \delta\left[E - H(\vec{p},\vec{q})\right]\mathrm{d}^N\vec{p}\,\mathrm{d}^N\vec{q}, \tag{1.5}$$

thus we find the normalization constant to

$$\omega(E) = \int \delta\left[E - H(\vec{p},\vec{q})\right]\mathrm{d}^N\vec{p}\,\mathrm{d}^N\vec{q}. \tag{1.6}$$

The normalization constant $\omega(E)$ is known as the *microcanonical partition function*. Looking at the definition of the microcanonical density distribution function and partition function we notice they have the dimensions of positions and momenta, thus rendering Eq. 1.4 not being a true probability density. This problem strictly arises due to the transformation of a discrete to a continuous probability density function, as it would be characteristic for quantum and classical mechanics, respectively. Consequently, to address the problem of dimensions we seek to find a constant, that can remove the dimensions in the classical regime, while simultaneously ensuring correct quantum to classical mechanical transition. The approach in doing so will be via determining the number of states for an ideal gas in the classical regime in which the Hamiltonian is chosen to only possess kinetic energy, and in the quantum regime in which we solve the Schrödinger equation. This derivation can be found in the end of this chapter. The resulting constant can be found to be the Planck constant to the power of $3N$. We will dub this constant the *fundamental volume* or the *quantum volume*, due to this volume being the smallest volume in which a microstate can be defined as equivalent to Heisenberg's uncertainty principle. This differential element of microstates, $\mathrm{d}\Gamma$, is then given by

$$\mathrm{d}\Gamma = \frac{1}{N!}\frac{\mathrm{d}^N\vec{p}\,\mathrm{d}^N\vec{q}}{h^{3N}}. \tag{1.7}$$

The appearance of the factor $1/N!$ in Eq. 1.7 is related to yet another quantum to classical mechanical issue. While different particles are distinguishable, even when they belong to the same specie in the classical regime, quantum mechanics reveals particles belonging to the same species to in fact be completely indistinguishable. The possibility to label particles yields more possible state in the classical regime over quantum mechanics, hence we correct for *over-counting*. There must be $N!$ ways of arranging $N$ particles, consequently we reduce the (classical) phase space volume by the factor $1/N!$. Returning to the microcanonical partition function; rewriting Eq. 1.6 in terms of $\mathrm{d}\Gamma$ we obtain

$$\omega(E) = \int_\Gamma \delta\left[E - H(\vec{p},\vec{q})\right]\mathrm{d}\Gamma. \tag{1.8}$$

In the calculation of many thermodynamic properties using the partition function, the fundamental volume constant is of no importance, due to its disappearance in differences for thermodynamic properties. However, it does aid in yielding a physical and philosophical understanding of the partition function: Despite the momenta and position coordinate are continuous, there are not an infinite amount of microstates in a finite enclosed volume, unlike there are for example an infinite amount of real numbers between the integers zero and one. Finally, as we are about to see, it shall also aid us greatly in the understanding of entropy, which is explicitly related to the partition function.

## 1.5 Entropy

With the possibility to count the number of states in a system of constant energy, we shall now attack the concept of *entropy*, which is the most complicated and mysterious classical property.[1] The complications arise due to the many physical interpretations applicable to entropy, some of which are related to the study of steam engines and others related to atomic theory. Our starting point for the discussion of entropy will be the famous Boltzmann entropy formula

$$S = k_B \ln[\omega(E)]. \tag{1.9}$$

Where $\omega(E)$ is the microcanonical partition function for a system of energy $E$ at constant volume and number of particles. In older literature $\omega(E)$ is also commonly donated the weight of the system, due to the strong connection to probability theory. Because the entropy simply being proportional to the logarithm of the number of available microstates, the difference in entropy between two macrostates is proportional to the logarithmic ratio of the number of microstates for the two macrostates. The driving force associated with the transformation of a microcanonical system from one macrostate into another macrostate is the *entropic force*. Since entropy is associated with probability, the entropic force is stochastic. This implies that while the direction of the entropic force is deterministic, the system's path towards equilibrium is usually not monotonic unless the system is found in the thermodynamic limit.[7] As such the entropic force is unlike classical forces, not a physical force. An example of this could be the expansion of a gas whose only interactions are via collisions: Increasing the volume occupiable by the particles would always lead towards an increase in entropy, due to more configurational microstates becoming accessible and is, therefore, a more probable state. In general, the entropy is indicating the direction towards the state of highest probability and is the most fundamental state function in determining spontaneous processes as it was discovered by Rudolf Clausius.[24]

In the previous paragraph, we discussed entropy on the level of changing the system in terms of thermodynamics variables such as volume, number of particles, or energy. However, it

[1] At least this is the opinion of the author.

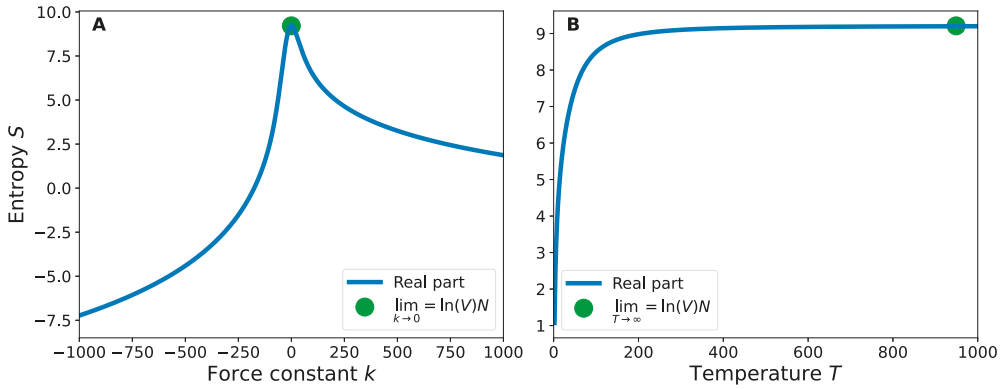



**Figure 1.1:** Configurational entropy of a harmonic oscillator ($U = 0.5kx^2$) as a function of **A** the force constant determining the oscillator strength of the harmonic oscillator and **B** the temperature given a constant force constant. For varying force constant, we see the entropy reaches a maximum for the force constant equal to zero with the entropy matching the entropy of an ideal gas ($S \propto \ln(V)N$), due to the system being completely uncorrelated. Introduction of attraction $k > 0$ or repulsion $k < 0$ causes the entropy to decrease due to the introduction of correlations. Similarly, increasing the temperature given a constant force constant for the harmonic oscillator will cause the system to behave more and more as an ideal gas, due to the thermal energy diminishing the effect of interactions, thus effectively removing correlations introduced in the system by the harmonic potential. In the calculation of the entropy for a harmonic oscillator complex number appears, however with the imaginary part canceling to zero in the final expression, leaving only the real part visualized in the plot.

remains questionable how the effect of interactions and in turn how "physical forces" affect the system. After all, they do have an impact on how systems transform on an everyday level. However, it turns out entropy already partly includes the effects of interactions. Introduction of interactions in a system creates correlations between particles, and these correlations reduce the number of accessible states. we shall therefore make the following conjecture: *For any given ensemble and choice of thermodynamic variables, the upper limit of entropy is always the ideal gas*. To illustrate this idea, the entropy for a harmonic oscillator has been visualized as a function of the force constant, which determines the strength of the interaction with increasing displacement from the rest position, in Fig. 1.1**A**. As expected given our conjecture, the entropy is at a finite maximum given the force constant is zero removing all interactions present in the system thus making the oscillator behaving like an ideal gas. In contrast, we find that both attraction ($k$ being positive) and repulsion ($k$ being negative) yield lower entropy. From this we see the direct link between disorder and entropy; a system with interactions imposes correlations that yields higher order, and in turn lower entropy.

## 1.6 The Isothermal Ensembles

Up to this point, we have only discussed systems of constant energy, which would be characteristic for isolated systems, that can not exchange energy or matter with the surroundings.

For a system of a constant number of particles, volume, and temperature, which we shall donate the *canonical (NVT) ensemble*, enclosed as a sub-volume of a microcanonical system, one may solve for the density distribution function from Liouville's theorem under the equilibrium condition (Eq. 1.3) and the constraint of maximum entropy. The resulting density distribution function is the famous Boltzmann distribution function

$$\rho_{NVT}(\vec{p},\vec{q}) = \frac{e^{-\beta H(\vec{p},\vec{q})}}{Q(T)}. \tag{1.10}$$

Here $\beta$ is the *thermodynamic* $\beta$ having the identity $\beta = \frac{1}{k_{\mathrm{B}}T}$, and $Q(T)$ is the *canonical partition function* having the same properties as the microcanonical partition function of ensuring correct normalization of the density distribution function

$$Q(T) = \int_{\Gamma} e^{-\beta H(\vec{p},\vec{q})}\, \mathrm{d}\Gamma. \tag{1.11}$$

In the previous equation the right-hand side is integrated over the phase space donated $\Gamma$, and $\mathrm{d}\Gamma$ is the fundamental volume used to distinguish between microstates in phase space as it was discussed in chapter 1.4. So far we have dealt with systems of constant volume, but with either constant energy as characteristic for an isolated system or constant temperature as it is characteristic for a closed system. Let's instead now look at an isothermal system with variable volume: the so-called *isothermal-isobaric* $(NPT)$ *ensemble*. An obvious derivation from Liouville's theorem is no longer simple without explicitly expressing the volume dependency of the Hamiltonian, such that the equations of motion will allow volume fluctuations. However, it is possible to derive under the assumption of the isothermal-isobaric ensemble being a sub-volume of a microcanonical system. The density distribution function for the $NPT$ ensemble is given by

$$\rho_{NPT}(\vec{p},\vec{q},V) = \frac{1}{\Delta(T,p)} e^{-\beta H(\vec{p},\vec{q},V)+pV}, \tag{1.12}$$

where $p$ is the pressure of the surrounding volume reservoir and $\Delta(T,p)$ is the *isothermal-isobaric partition function*:

$$\Delta(T,p) = \int_{\Gamma}\int_{V=0}^{V=\infty} e^{-\beta H(\vec{p},\vec{q},V)+pV}\, \mathrm{d}V\, \mathrm{d}\Gamma. \tag{1.13}$$

Where the first integral occurs over the phase space and the second integral is taken over the possible system volumes. From the isothermal-isobaric partition function, we see an interesting feature that leads us back to the physics of counting states: The canonical ensemble is a subset of the isothermal-isobaric ensemble, due to the first integral being equal to the canonical partition function, thus meaning the isothermal-isobaric ensemble can be thought of as the canonical ensemble with one additional degree of freedom being the

volume. Similarly, one might also say that the microcanonical ensemble is a subset of the canonical ensemble due to the equivalent integration over various energy levels, with the probability of the individual energy layers determined by the temperature. However, any average obtained in any ensemble is the same in the thermodynamic limit ($N \to \infty$) as energy fluctuation in the canonical ensemble, and volume fluctuations in the isothermal-isobaric ensemble cease to exists rendering all ensembles equivalent to the microcanonical ensemble.

Returning to the entropy; with the statement of systems in different ensembles being subsets of one another, one can anticipate the entropy for finite systems not found in the thermodynamic limit, and the different ensembles must possess different entropies. As an example, if one were to do a simulation in the microcanonical and determine the average temperature for the system, and now do a simulation in the canonical ensemble using the determined temperature from the microcanonical ensemble, the entropy in the canonical could be thought of as a sum of the entropies for the various energy layers, thus causing $S_{\text{canonical}} \geq S_{\text{microcanonical}}$. Gibbs, however, discovered the ensemble dependency could be abolished using the mean Boltzmann entropy, in which the entropy is averaged overall energy levels

$$S = \langle S(E_i) \rangle = \int p(E_i) S(E_i) = -k_{\text{B}} \int p(E_i) \ln[p(E_i)] = -k_{\text{B}} \langle \ln[p(E_i)] \rangle. \quad (1.14)$$

## 1.7 Focusing on the System: Free Energy

*The Danish saying: "Don't jump over the fence where it is lowest" versus my father: "Only an idiot jumps over the fence where it is the highest"*
— If my father were to invent the concept of free energy.

We established that entropy is the variable in determining the natural direction of systems in the microcanonical ensemble. With the introduction of the canonical ensemble and the isothermal-isobaric ensemble we stated these ensembles to be a sub-volume of the microcanonical ensemble, as a consequence to determine the natural direction of these ensembles, we need to do the annoying task of calculating both the change in entropy for the surroundings as well as the system,

$$\Delta S_{\text{total}} = \Delta S_{\text{surroundings}} + \Delta S_{\text{system}}. \quad (1.15)$$

Within classical thermodynamics a system at constant volume and temperature can only do work via the heat process, which is equal to the change in internal energy of the system $\Delta U$, which in turn can be related to the entropy of the surroundings $\Delta S_{\text{surroundings}} =$

$-\Delta U/T$. We can now rewrite the total entropy as follows

$$-T\Delta S_{\text{total}} = \Delta U_{\text{system}} - T\Delta S_{\text{system}}. \tag{1.16}$$

We define the auxiliary function, which came to be known as the Helmholtz free energy $\Delta A = \Delta U - T\Delta S$. In a similar way for a system of constant pressure and temperature, we can define the Gibbs free energy $\Delta G = \Delta H - T\Delta S$. The gain of these free energy functions is the possibility to obtain the total entropy of the system and surroundings only given knowledge about the system, and thus allowing the surroundings to be completely ignored. Thus, the determination of free energies is an essential concept within statistical mechanics, as it allows to calculate the spontaneous direction of processes.

## 1.8 Connecting the Quantum & Classical Regime of Statistical Thermodynamics

In order to ensure the proper transition from quantum to classical mechanics, we choose to count the number of states for an ideal gas. To do so lets first look at a rewriting of the microcanonical partition function. In Eq. 1.4 we require the energy to be fixed, however, if we are instead to take the energy to lie within a small accepted range donated $[E - \Delta, E]$, we can rewrite the microcanonical density distribution function as

$$\lim_{\Delta\to 0} \rho^{\Delta}_{NVE}(\vec{p},\vec{q}) = \lim_{\Delta\to 0} \frac{\Delta\delta\left[E - H(\vec{p},\vec{q})\right]}{\Delta\omega(E)} = \rho_{NVE}(\vec{p},\vec{q}). \tag{1.17}$$

As we can see, in the limit of the energy decrement $\Delta$ approaching zero, we recover the microcanonical density distribution function. Eq. 1.17 will become useful for the upcoming derivation.

### The Quantum Mechanical Scenario

As with almost all of quantum mechanics our starting point will be the Schrödinger equation

$$-\frac{\hbar^2}{2m}\Delta\psi_i(\vec{x}) = E_i\psi_i(\vec{x}). \tag{1.18}$$

We choose the gas should be found in a rectangular box with the lengths $a_1$, $a_2$, and $a_3$ and periodic boundary conditions so we have the constraint

$$\psi(\vec{x_1},\vec{x_2},\vec{x_3}) = \psi_i(\vec{x_1}+a_1,\vec{x_2},\vec{x_3}) = \psi_i(\vec{x_1},\vec{x_2}+a_2,\vec{x_3}) = \psi_i(\vec{x_1},\vec{x_2},\vec{x_3}+a_3). \tag{1.19}$$

The solution to the Schrödinger equation given this situation is

$$\psi_i(\vec{x}) = \frac{1}{\sqrt{V}} \exp\left(\frac{i}{\hbar}\vec{p}\,\vec{x}\right), \tag{1.20}$$

where $V$ is the volume and $\vec{p_i}$ is given by

$$\vec{p_i} \in \left\{ 2\pi\hbar \left(\frac{n_1}{a_1}, \frac{n_2}{a_2}, \frac{n_3}{a_3}\right) \middle| \, n_i \in \mathbb{Z} \right\}. \tag{1.21}$$

The wave functions of the energy eigenstates given $N$ particles is then given as the product of the individual particle wave functions

$$\psi(\vec{x}) = \prod_{i=1}^{N} \psi_i(\vec{x}) = \frac{1}{\sqrt{V}} \prod_{i=1}^{N} \exp\left(\frac{i}{\hbar}\vec{p}\,\vec{x}\right). \tag{1.22}$$

Due to the principal quantum numbers, $n$, can only take the values of integers, the momenta can be viewed as a lattice in a $3N$-dimensional momentum space with a corresponding density of states being

$$\rho = \frac{1}{\left[\left(\frac{2\pi\hbar}{a_1}\right)\left(\frac{2\pi\hbar}{a_2}\right)\left(\frac{2\pi\hbar}{a_3}\right)\right]^N} = \frac{1}{\frac{(2\pi\hbar)^{3N}}{V^N}}. \tag{1.23}$$

From the partition function on the form presented in Eq. 1.17 we need to calculate the number of lattice points within the range $\sqrt{2m(E-\Delta)}$ and $\sqrt{2mE}$. In the limit of a large box we get $2\pi\hbar/a \ll 1$ rendering the error of assuming continuous probability density small, and thus the number of states can be obtained by multiplying with the volume

$$\begin{aligned} \omega(E) &= \frac{V^N}{(2\pi\hbar)^{3N}} \left[ V_{3N}\left(\sqrt{2mE}\right) - V_{3N}\left(\sqrt{2m(E-\Delta)}\right)\right] \\ &= \frac{1}{\frac{3N}{2}\Gamma\left(\frac{3N}{2}\right)} \left[ V \left(\frac{1}{2\pi\hbar^2}\right)^{\frac{3}{2}} \right]^N \left\{ (2mE)^{\frac{3N}{2}} - (2m[E-\Delta])^{\frac{3N}{2}} \right\}. \end{aligned} \tag{1.24}$$

## The Classical Mechanical Scenario

The Hamiltonian of an ideal gas containing $N$ particles of identical mass $m$ is given by merely the kinetic energy

$$H(\vec{p}, \vec{q}) = \sum_{i=1}^{N} \frac{\vec{p_i}^2}{2m}. \tag{1.25}$$

Using the microcanonical partition function on the form presented in Eq. 1.6, we obtain

$$\omega(E) = \int \delta\left(E - \sum_{i=1}^{N} \frac{\vec{p}_i^{\,2}}{2m}\right) \mathrm{d}^{3N}p\, \mathrm{d}^{3N}q. \tag{1.26}$$

In the above expression we need to evaluate the integration over positions and momenta. For the positions we find the Hamiltonian does not depend on the positions of the particles, $\vec{q}$, and hence the integration over positions can be found to yield the volume to the power of $N$ particles. To ease the integration over momenta we rewrite the microcanonical partition function in terms of spherical coordinates

$$\begin{aligned}\omega(E) =& V^N \int \mathrm{d}\Omega \int_0^\infty |\vec{p}|^{3N-1} \delta\left(E - \frac{|\vec{p}|^2}{2m}\right) \mathrm{d}|p| \\ =& V^N \int \mathrm{d}\Omega \int_0^\infty |\vec{p}|^{3N-1} \frac{m}{\sqrt{2mE}} \times \\ & \left[\delta\left(\sqrt{(2mE)} - |\vec{p}|\right) + \delta\left(-\sqrt{(2mE)} - |\vec{p}|\right)\right] \mathrm{d}|p|.\end{aligned} \tag{1.27}$$

Where the last equality uses a Dirac $\delta$-function identity for functions.[2] Due to the integration from zero to infinity, the second term of the square brackets can never be anything but zero, and hence vanishes from the expression. Furthermore, recalling the Heaviside step function ($\Theta$) to be the derivative of the Dirac $\delta$-function we find

$$\begin{aligned}\omega(E) &= V^N \frac{m}{\sqrt{2mE}} \frac{\mathrm{d}}{\mathrm{d}\sqrt{2mE}} \left(\int \mathrm{d}\Omega \int_0^\infty |\vec{p}|^{3N-1} \Theta\left(\sqrt{2mE} - |\vec{p}|\right) \mathrm{d}|\vec{p}|\right) \\ &= V^N \frac{m}{\sqrt{2mE}} \frac{\mathrm{d}V_{3N}\left(\sqrt{2mE}\right)}{\mathrm{d}\sqrt{2mE}} = V^N \frac{m}{\sqrt{2mE}} A_{3N}\left(\sqrt{2mE}\right).\end{aligned} \tag{1.28}$$

In the above expression $V_{3N}$ and $A_{3N}$ are the volume and surface area of a $3N$-dimensional sphere with radius $\sqrt{2mE}$. Substituting the expression for a $3N$-dimensional surface area we obtain

$$\omega(E) = \frac{1}{\Gamma\left(\frac{3N}{2}\right)} \frac{1}{E} \left[V\left(2\pi E m\right)^{\frac{3}{2}}\right]^N. \tag{1.29}$$

Up til this point we have done the derivation under the assumption of fixed energy $E$. Allowing the minor energy fluctuations in the range $[E-\Delta, E]$ and using the corresponding microcanonical partition function (Eq. 1.17) using the exact same steps as previously we obtain

$$\begin{aligned}\omega_\Delta(E) &= V_{3N}\left(\sqrt{2mE}\right) - V_{3N}\left(\sqrt{2m(E-\Delta)}\right) \\ &= \frac{1}{\frac{3N}{2}\Gamma\left(\frac{3N}{2}\right)} \left[V\left(2\pi\right)^{\frac{3}{2}}\right]^N \times \left[\left(2mE\right)^{\frac{3N}{2}} - \left(2m[E-\Delta]\right)^{\frac{3N}{2}}\right].\end{aligned} \tag{1.30}$$

[2] $\delta(f(x)) = \frac{1}{|f'(x)|} \sum_{f(x_i)=0} \delta(x_i - x)$

## Unifying the Quantum & Classical Mechanical Scenarios

To compare the number of states given the classical methodology and quantum methodology we take the ratio of the partition functions determined by the individual methods. The ratio of the quantum (Eq. 1.24) and classical (Eq. 1.30) microcanonical partition function is given by

$$\begin{aligned}\frac{\omega_{\Delta}^{\mathrm{QM}}(E)}{\omega_{\Delta}^{\mathrm{CM}}(E)} &= \frac{\frac{1}{\frac{3N}{2}\Gamma\left(\frac{3N}{2}\right)}\left[V\left(\frac{1}{2\pi\hbar^2}\right)^{\frac{3}{2}}\right]^N \times \left\{(2mE)^{\frac{3N}{2}} - (2m[E-\Delta])^{\frac{3N}{2}}\right\}}{\frac{1}{\frac{3N}{2}\Gamma\left(\frac{3N}{2}\right)}\left[V\left(2\pi\right)^{\frac{3}{2}}\right]^N \times \left\{(2mE)^{\frac{3N}{2}} - (2m[E-\Delta])^{\frac{3N}{2}}\right\}} \\ &= \frac{1}{h^{3N}}.\end{aligned} \tag{1.31}$$

From Eq. 1.31 it is now clear that any other choice than the Planck constant to the power of the dimensionality times the number of particles $N$ of the system in the expression for the fundamental volume (Eq. 1.7) would yield an inconsistency between quantum and classical mechanics in the number of states for an ideal gas, and hence the entropy and likewise the free energy would also be off, as these quantities are directly related to the partition function. This derivation relied in particular on one approximation found in the quantum mechanical derivation, which is the continuity of energies in the limit of a large box. This has the implication, that the translational contribution to the molecular partition function (being the only contribution to a monatomic ideal gas) can be written as the volume of the system over the Thermal de Broglie wavelength ($\lambda \equiv \sqrt{2\pi\hbar\beta/m}$) cubed. Hence we find classical Maxwell-Boltzmann statistics is a good approximation when $\lambda^3/V \ll 1$, otherwise quantum statistics such as Bose-Einstein statistics or Fermi-Dirac statistics has to be applied.

# Chapter 2

# Molecules of Life

*When we have broken down living systems to molecules and ... analyzed their behavior, we may kid ourselves into thinking that we know what life is, forgetting that molecules have no life at all.*
— Albert Szent-Györgyi

## 2.1 Water

In the exploration of our universe for the search of life, water is of great importance due to it being one of the essential substances required to maintain life as we know it. On our planet, Earth, water alone constitutes 71 percent of the planet's surface, thus making it the most abundant liquid. The impact of water on Earth is remarkable, influencing the landscape (e.g. the formation of Grand Canyon by water erosion), climates (e.g. the dry desserts of Sahara to the ever rainy Sweden), and finally the biology of organisms, due to it being the most abundant solvent. The physical properties of water are remarkable with water being one of the few substances to expand upon freezing (maximum density in liquid phase), resulting in water freezing first at the air-water interface and then downwards, thus enabling life within the deep sea and for all aquatic life not to die every winter. Another remarkable property of water is its large heat capacity, being in simple terms a measure of how much energy a material can store without increasing the temperature of the system, thus enabling water to act as a thermostat in regulating the temperature of Earth. Finally, the dielectric properties of water are also worth mentioning, possessing a static dielectric constant of 78.4 at 25 °C having the implication electrostatic interactions between ions on distances longer than 7.2 Å are energetically comparable to the thermal energy. This final property has the significance that salts can be commonly dissolved in water. Water has

almost a university solvent action, nearly all chemicals can be dissolved in water to some extent, even small factions of oil. As such water is truly a highly corrosive chemical, yet considered physiologically harmless.

Due to the highly interesting properties of water and its abundance, the modeling of water has been a greater scientific challenge. In particular, it has proved very difficult to construct a water model which can reproduce thermodynamic experimental properties of bulk water, such as dielectric constant, critical points of phase transitions, and heat capacities, but also dynamic properties such as self-diffusion and geometrical properties as found from quantum mechanics. Among the most commonly utilized water models belong the "simple point charge" (SPC) family and the "transferable intermolecular potential" (TIP) family. While the SPC family all utilize so-called "3 points" models, the TIP family is containing multiple point models ranging between 3 and 5 points to achieve the correct tetrahedron geometry of water. Comparing the SPC family and the TIP 3 point (TIP3P) model, the most outstanding is the SPC and extended SPC (SPC/E) ability to reproducing the self-diffusion of water and other bulk properties of water. On the other hand, TIP3P usually performs well-reproducing solvation properties, while failing in reproducing bulk properties. The field of constructing better and better water models are constantly evolving and hence it is difficult to give a comprehensive and detailed review that also reflects today's knowledge.

## 2.2 Proteins

Proteins are essentially found everywhere in biological systems, being the main workhorse molecule of life. Examples of the purposes of proteins include enzymes responsible for the separation and joining of molecules of life, virus capsids and antibodies and their eternal battle in the bloodstream, photosynthesis, energy production, storage, infrastructure and structure, and finally the creation of more proteins. A great variety exists due to the 20 fundamental standard proteinogenic amino acids. The amino acids are characterized by possessing an amino and a carboxyl functional group, constituting the protein *backbone*, and a unique side chain for each of the 20 amino acids. The uniqueness of the side chain is the main contribution to the heterogeneity of proteins, and thus the side chain is usually used to characterize and categorize the amino acids. The amino acids are most commonly categorized into the three categories; hydrophobic, polar, and charged, however many other categorization schemes based on the amino acid's properties are possible with examples including size, polarity, hydropathy, and disorder promoting. Consequently, a small protein with 50 amino acids (like insulin) could generate more than $10^{65}$ sequences, also known as the primary structure, many of which would have different properties. For proteins to execute their specific functions, they usually fold into specific three-dimensional structures, characterized by the formation of highly ordered structural elements such as $\alpha$-helices and

$\beta$-sheets, known second secondary structure, with the three-dimensional arrangement of the elements known as the tertiary structure.

The previously mentioned folded state of proteins is usually termed to be the native state, due to the folded state being mostly populated at native conditions, while the protein in the unfolded state is termed the denatured state, due to the unfolded state being mostly populated at denaturing conditions. However this is only one example of the many transitions proteins can undergo, other possible transitions include aggregation, crystallization, fibrillation, misfolding, and phase separation. All of these states are commonly thermodynamically stable, with the transition between the states typically being reversible. Caution however needs to be exercised as some of the transformations are irreversible. If the change conducted appear longer than on the experimental timescale, as it would, for example, be characteristic if changes in covalent bonding are introduced to the protein by for example high-temperature perturbation. Another example that does not involve changes in covalent bonding, could be the formation of a highly entangled and stabilized intermediate state stabilized by non-covalent bonding thus heavily decelerating the formation of the correctly folded state.

The mentioned equilibria of proteins can all be perturbed by physicochemical parameters such as temperature, pressure, pH, ionic strength, and co-solvent but also by amino acid residue substitutions. While it is hard to predict the exact numeric effect of the perturbation, and sometimes even the qualitative effect, it is possible to illuminate many thermodynamic properties of the structural stability of well-behaving proteins using simple solvation thermodynamics and lattice statistics.

### 2.2.1 Structural Stability Perturbations Illuminated by a Simple Transfer Model

To reduce the complexity of proteins, consider each amino acids a sphere located in a lattice, connected by stiff bonds. We define the native, folded state as a single conformation, characterized by a high amount of stabilizing interactions.[29,30] The entropy contribution to the folding of the chain, assuming it to be distinguishable and independent, is given by[131]

$$\Delta S_{\text{fold}} = S_{\text{N}} - S_{\text{D}} = R \ln \left( \frac{Q_{\text{N}}}{Q_{\text{D}}} \right) = RN \ln z, \tag{2.1}$$

where we have approximated the canonical partition function of the native state to be one, equivalent to the structure being completely static, and $z$ is the molecular partition function. The interactions introduced in the native state, we can model as the difference in solvation free energy (fixed position solvation process, cf. chapter 3.1) of the individual amino acids. Rewriting Eq. 2.1 in terms of the free energy we obtain

$$\Delta G_{\text{fold}} = N(g + RT \ln z) \tag{2.2}$$

where $g$ is free energy of transfer of an amino acid residue into a non-aqueous protein environment. Given this simple model the energetics of folding are thus governed by the competition of the solvation of the amino acid residues in the native state of the protein and the conformational entropy of the denatured state.[29,30,152]

The protein stability is well-known to be highly dependent on the presence of denaturing or stabilizing co-solutes. Among the most common denaturing co-solutes are urea and guanidine hydrochloride, where as stabilizing co-solutes includes trimethylamine N-oxide (TMAO) or L-arginine L-glutamate salt. It was first demonstrated by Tanford[130] and later by Green and Pace,[43] and by Santoro and Bolen,[115] the free energy of folding is linearly dependent on the co-solute concentration for many proteins. Within the simplified transfer model, as first illustrated by Tanford[131] the transfer free energy of the amino acid residues can be written as

$$g(c) = g_0 + m_i c, \tag{2.3}$$

where $g_0$ is the transfer free energy of the residue in pure water, and $m_i$ is the residual $m$-value. Substituting into the expression for the free energy of folding (Eq. 2.2), we obtain

$$\Delta G_{\text{fold}} = N(RT \ln z + g_0 + m_i c) \tag{2.4}$$

Where the $m$-value is defined as $Nm_i$. From this simple model, it now becomes clear the $m$-value should be related to the degree of newly exposed surface area upon folding, which is also proportional to the number of residues $N$ in the protein. This was experimentally found to be true.[86] Despite the great number of discoveries experimentally, the strength of the simple transfer model, the molecular mechanism responsible for the denaturation of proteins is still a topic up for discussion. In particular, it is still discussed if the perturbation of the native state is due to an indirect mechanism of action, in which co-solute interact with water, and thus weakening the interactions between protein and water or the direct mechanism of action, in which co-solute interact with the protein, protecting or exposing molecular group sensitive to hydrophobic interactions. Another possibility yet to be explored is the possibility of the denatured state stabilization, which is difficult to illuminate by experimental and computational methods.

# Chapter 3

# Solvation Thermodynamics

*I am telling you how difficult a "why" -question is. You have to know what it is that you are permitted to understand and allowed to be understood and known, and what it is you are not. You notice... the more I ask "why?", it gets interesting. That is my point: the deeper it is the more interesting it gets.*

— Richard Feynman in an interview.

With the topic of *solvation thermodynamics* the terms *solubility* and *solvation* are central. While *solubility* refers to a quantitative measure of a solute's preferential occupancy in a given solvent, *solvation* refers to the process of inserting the solute molecule into a specific solvent. Given these definitions, the solubility is very broadly defined to include all measures yielding insight into the solute's preference for one phase over another. On the experimental side, the perhaps most common phase-equilibrium would be the equilibrium for the solute to be in various states of matter, for example, the formation of a precipitate from the solution, upon reaching the saturation limit. We could write such equilibrium as

$$\mathrm{Solute}_{(s)} \rightleftharpoons \mathrm{Solute}_{(aq)}, \quad K = \frac{\left[\mathrm{Solute}_{(aq)}\right]}{\left[\mathrm{Solute}_{(s)}\right]}, \tag{3.1}$$

where $K$ is the solubility equilibrium constant. Another example within the category of phase-equilibrium is the vapor-liquid equilibrium which addresses the preference of the solute to form either a liquid or a gas. We can qualitatively understand the equilibrium by considering any isothermal closed molecular system will be having a Boltzmann distributed energies among the molecules, thus a specific faction will possess so much kinetic energy they can escape the otherwise attractive interaction formed in the liquid phase, entering the gas phase. Consequently, the vapor pressure formed depends on the temperature and the intermolecular interactions formed in the liquid. Given this knowledge, we can rationalize

that water forms stronger intermolecular interactions than ethanol which forms stronger interactions than acetaldehyde, due to the lower vapor of water compared to ethanol, and ethanol having a lower vapor pressure compared to acetaldehyde.

Another category of solubility equilibria is solute partitioning between an organic and water phase, which reveals the contrast in stabilizing interactions, most commonly in terms of hydrophobic and hydrophilic interactions. The last category, which is perhaps the most experimentally inconvenient one, is vacuum-solvent the equilibrium. This equilibrium reports directly the work exerted by the specific composition solvent on the solute.

The different solubility measures previously mentioned are all related to investigating the interactions between solvent on the solute, with each method reporting something unique about the properties of the solute. As such, we are now able to specify the focus of the topic solvation thermodynamics, namely the quantification and qualification of the work necessary for inserting molecular matter into a given solvent. Consequently, we will in this chapter discuss the solvation process, which identifies the various challenges associated with the insertion of solute into a solvent. Additionally, we will establish the statistical thermodynamics framework necessary to relate macroscopic observables to microscopic events, and finally introduce computational free energy calculations methods to predict the solubility of molecular matter.

## 3.1 The Solvation Process

As previously stated, the term *solvation* refers to the process of inserting the solute molecule into a specific solvent. To determine if the solvation of solute is a spontaneous process, the difference in free energy before and after the insertion must be negative, which can be decomposed into multiple contributions of enthalpic[1] and entropic nature. An example of the solvation process is illustrated in Fig. 3.1. In this specific example, we see the transfer of a chloride anion (solute) from vacuum to water (solvent), i.e.the hydration of a chloride anion. In the specific case where water is the solvent, one commonly uses the term *hydration*, which is more specific than solvation. In the first step of Fig. 3.1 we see the creation of a cavity. The formation of a cavity is both an enthalpic and an entropic unfavourable process. The enthalpic contribution arises from the breaking of, usually stabilizing, intermolecular bonds, which in the case of water would predominantly be hydrogen bonds. The entropic contribution appears due to the excluded volume, created for the solute, thus reducing the accessible configurational phase space volume for water. The free energy of this step is strongly dependent on the size of the solute, with larger solutes requiring more

[1] We will generalize and use the term enthalpy/enthalpic regardless of the ensemble, this is due to the enthalpy is defined as $\Delta H = \Delta U + \Delta(pV)$. Therefore, if no $pV$ work is done on or by the system the enthalpy and internal energy are equivalent, $\Delta H = \Delta U$.

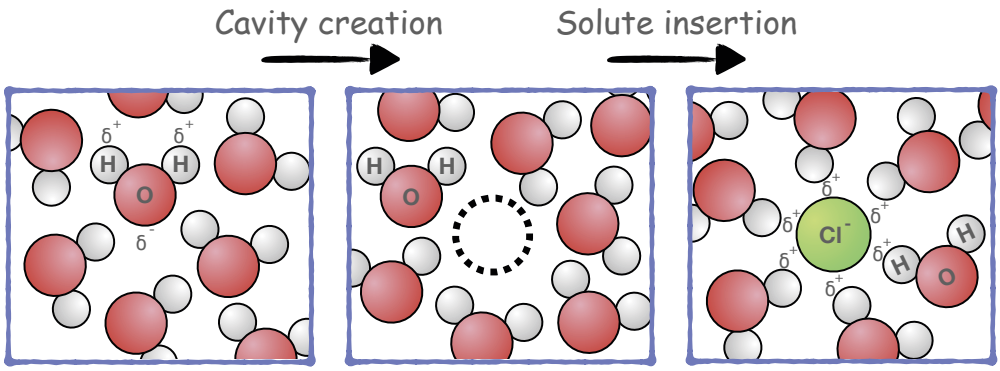



**Figure 3.1:** Illustration of the solvation process in which the solvent (water), is creating a cavity equal to the excluded volume of the solute (a chloride anion), characterized by breaking intermolecular interactions between solvent molecules, in this case, hydrogen bonds. In the last step where the solute is finally inserted the solvent is re-orientating to a preferential configuration corresponding to intermolecular interactions with low free energy between solvent-solute, solvent-solvent, and solute-solute.

ordering of solvent by the created cavity followed by a larger number of solvent-solvent interactions broken. As a result, the choice of solvent is also important and in particular the strength of solvent-solvent interactions. The second step of Fig. 3.1, is to turn on the possible interactions to the solute. So far we have already established the hydration is unfavorable, so to turn the whole process favorable, the solute-solvent interactions must be highly stabilizing. In the specific case for the hydration of chloride, strong electrostatic interactions can be formed between the negatively charged anion and the positive partially charged hydrogen of water yielding an enthalpically favorable contribution to the hydration of the anion, while the preferential re-orientation of the solvent around the solute, yielding an entropically unfavorable contribution.

The last stage of the hydration of the chloride anion is to free the anion. So far we have assumed the particle to be grown into a fixed position in the system, thus an entropic favourable contribution of mixing is achieved by releasing the particle from its position of growth, allowing it to diffuse in the system. Finally, there is the possibility for ensemble effects, which are particularly linked to the re-organization energies. For a constant-pressure solvation process, an associated relaxation of the system volume can be accompanied with the solvent re-organization of solvation, which can be up to several $kT$ different compared to solvent re-organization associated with a constant-volume solvation process.

From the above, it should now be evident the solvation of molecular matter is no trivial task, with many contributions driving the process in different directions. However keeping these contributions in mind, we can construct a statistical thermodynamic description to estimate the various contributions. Thankfully, the frontiers in the field of solvation thermodynamics have already craved a great deal of the way to mathematically describe the processes.

## 3.2 The Chemical Potential of Solute

As we saw from the solvation process, we are looking at the equilibrium of transferring a particle into a solvent ($N \rightleftharpoons N+1$), the chemical potential in the canonical ensemble of an infinitely dilute solute in the thermodynamic limit, is

$$\begin{aligned}\mu_{\text{solute}} &= \left[\left(\frac{\partial A}{\partial N_s}\right)_{N_{\text{solvent}},V,T}\right]_{N_{\text{solute}}\to 0} \\ &= A(N_{\text{solvent}}, N_{\text{solute}}=1, V, T) - A(N_{\text{solvent}}, N_{\text{solute}}=0, V, T),\end{aligned} \tag{3.2}$$

where $A$ is the Helmholtz free energy of the system containing $N_{\text{solvent}}$ solvent and $N_{\text{solute}}$ solute particles enclosed in volume $V$ at temperature $T$. Furthermore, we linearly decompose the chemical potential into an ideal contribution $\mu^{\text{id}}$ and an excess contribution $\mu^{\text{ex}}$

$$\mu_{\text{solute}} = \mu^{\text{id}} + \mu^{\text{ex}}. \tag{3.3}$$

The ideal contribution is the chemical potential of an ideal gas at the stated conditions, thus all intermolecular interactions are omitted from the system and thus the contribution arises from the kinetic energy and the possibility to occupy the enclosed volume. The excess chemical potential is therefore related to the correlation arising from the intermolecular interactions, which can be expressed through the canonical configurational integrals

$$\beta\mu^{\text{ex}} = -\ln\left[\frac{Z(N_{\text{solute}}=1)}{Z(N_{\text{solute}}=0)}\right]. \tag{3.4}$$

Eq. 3.4 is central for the calculation of free energy calculations and will be considered in much more detail in chapter 3.3. Instead, for now, we will focus on the ideal contribution to the solvation free energy.

### 3.2.1 The Chemical Potential of an Ideal Gas

The perhaps easiest model particle system is an ideal gas. As it was seen in Eq. 3.3, we conveniently decompose the chemical potential of solute into the contribution arising from an ideal gas and an excess term arsing from the intermolecular interactions. Thus here we will investigate ideal gas contribution to the chemical potential. The starting point will be the rewriting of Eq. 3.2 to correspond to the chemical potential in the microcanonical (*NVE*) ensemble instead of the canonical (*NVT*) ensemble

$$\mu = \left(\frac{\partial A}{\partial N}\right)_{V,T} = -T\left(\frac{\partial S}{\partial N}\right)_{V,E} \tag{3.5}$$

the microcanonical chemical potential thus refers to the insertion of a particle such that the internal energy remains fixed. The entropy of a monatomic ideal gas can be obtained by inserting the microcanonical partition function for an ideal gas (Eq. 1.30) into the Boltzmann entropy formula (Eq. 1.9) and taking the Sterling approximation of the expression. The final result would be the famous Sackur-Tetrode equation (named after Otto Sackur and Hugo Tetrode who independently derived it in 1912)[44,110,111,135]

$$\frac{S}{k_B N} = \ln\left[\frac{V}{N}\left(\frac{4\pi m U}{3h^2 N}\right)\right] + \frac{5}{2}, \tag{3.6}$$

where the fraction $\frac{4\pi m}{3h^2}$ is the inverse fundamental volume per particle $v_Q$ as previously discussed in chapter 1.4 and 1.8. It should once more be noted that the Sackur-Tetrode equation is limited to the classical regime i.e. the volume of the system is greater than the fundamental volume ($V \gg N v_Q$) and thus can be described by Maxwell–Boltzmann statistics. Assuming the fundamental volume, $v_Q$, is constant, the difference in entropy upon a particle increment of one and thus the chemical potential yields

$$\mu = -T[S(N+1) - S(N)] = -T k_B \left(\ln\left[\frac{V}{N v_Q}\right] + \frac{3}{2}\right) \tag{3.7}$$

For the condition where the volume of the system is much greater than the total fundamental volume ($V \gg N v_Q$), we find the ideal chemical potential to always be negative, and thus the corresponding entropy of the system to be positive. The physical understanding of the ever negative chemical potential (positive entropy) of an ideal gas is somewhat trivial; more microstates with the same energy becomes accessible when more particles can be placed in the system.

Perhaps the easiest rationalization of this can be achieved by considering the distribution of particles in the configurational space of a system with constant volume. Imagine a 2x2 ensemble of distinguishable boxes which together constitute the volume of the system as illustrated in figure 3.2**A**. Each box represents the fundamental volume, meaning all quantum effects between particles are neglectable beyond this distance and no interactions between the particles are possible. As we can see from 3.2**A**, the number of microstates depends on the number of particles we are allowed to place within the system, with a maximum of 6 macrostates given 2 particles. At constant energy, the Boltzmann entropy formula ($S = k_B \ln W$) connects the entropy of the system's macrostate to the weight of the given macrostate, where the weight of the system is given by the number of microstates accessible. The number of permutations $W$ given a finite number of particles $N$ and finite number of accessible boxes $M$ yields

$$W(M, N) = \frac{M!}{N!(M-N)!}. \tag{3.8}$$

Eq. 3.8 is the well-known binomial coefficient and has the property of being an increasing function for $N \leq M/2$, which in the case of our particle system means the entropy upon

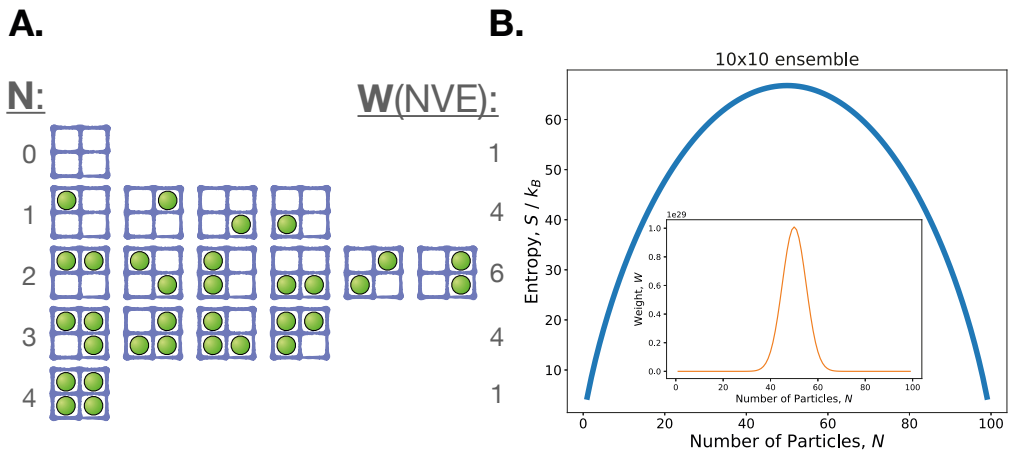



**Figure 3.2:** Illustration of the configurational contribution to the entropy upon insertion of ideal particles (chemical potential) into a system containing an ideal gas. **A.** Visualization of the weight of $N$ particles in a 2x2 ensemble at constant energy and volume (microcanonical), with the weight given by the binomial coefficient. **B.** Entropy as a function of the number of particles in a 10x10 ensemble at constant energy and volume, with the entropy calculated using Boltzmann's entropy formula. The insert is the corresponding weight of the system as a function of the number of particles.

inserting particles is increasing until the volume occupied by the particles exceeds more than half the system volume. This is illustrated in figure 3.2**B** in which the maximum entropy is achieved when the box is half occupied, with the entropy first increasing reaching a maximum at $N = M/2$ following a decrease in entropy. This conclusion is equivalent to the result of Eq. 3.7 and the necessary assumption the quantum volume occupied by the particles must be much smaller than the system volume, which is true in all practical cases.

The simple model presented above utilizing the binomial coefficient to calculate the ideal configurational chemical potential and entropy of the system due to quantum volume exclusion, could with interest be expanded to the study of hard-sphere systems, which is also a "hard" particle exclusion, to estimate the density at which the insertion of particles is no longer favorable.

## 3.3 Solvation Free Energy Calculations

To gain insight into the spontaneity of transformations of systems and the maximum amount of work the system can do on the surroundings, knowledge of the free energy is essential. Consequently, the development of methods to calculate free energies has drawn a lot of attention within the field of molecular simulations, with the field dubbed *free energy methods* or *free energy calculations*. The transformation of the system in question can vary greatly depending on interest and in theory one may conduct any transformation desirable.

Examples include changing the thermodynamic state of the system by varying the volume, pressure, or the number of particles.[2] However, it is also possible to calculate the free energy of chemical reactions, meaning to change the composition of atoms, the binding of ligands to host molecules, and the free energy of physical equilibria, meaning the transformation between thermostatically stable states of matter, with a specific example being the free energy difference between the folded and unfolded state of a protein. Within the field of solvation thermodynamics, we are, as earlier stated, interested in the transformation that involves moving solute from one environment to another. By this broad definition, there are a great number of specific processes which can be conducted each of which poses different challenges. In this chapter, we will thus look at methods aimed to calculate free energy changes for solutes into different solutions, but also methods directly aimed at determining the chemical potential, which we previously showed is a fundamental key concept of solvation thermodynamic.

The fundamental starting point is the utilization of the statistical mechanical expressions for free energy. In the canonical ensemble, the Helmholtz free energy $F$ for a system of $N$ distinguishable particles is given by

$$F(N,V,T) = -k_B T \ln Q = -k_B T \ln \left[ h^{-3N} \int_{\Gamma} e^{-\beta H(\vec{p},\vec{q})} \, \mathrm{d}\vec{p}\mathrm{d}\vec{q} \right], \tag{3.9}$$

where $\Gamma$ is the accessible phase space containing all the atomic degrees of freedom, namely the particle positions and momenta. Eq. 3.9 dictates the *absolute* free energy is given by the integration of the Boltzmann factor over the whole accessible phase space $\Gamma$, or in other words the free energy is proportional to the logarithmic canonical partition function. Eq. 3.9 simultaneously also reveals the essential difficulty in calculating the absolute free energy of a system, due to the free energy being dependent on a $6N$-dimensional integration to be conducted. Due to the partition function being an ever-positive function and the logarithm being a monotonically increasing function, the free energy will progressively become lower, as more and more regions of the phase space are included in the integration. Practically, this has the implication, that the free energy is a slow converging function for systems possessing large and complicated phase spaces, rendering it impossible to estimate the free energy accurately. Instead one estimates the difference in free energy between two states $i$ and $j$. Using Eq. 3.9 we find the difference in free energy to be

$$\begin{aligned} \Delta F_{i\to j}(N,V,T) &= F_j(N,V,T) - F_i(N,V,T) = -k_B T \ln \left[ \frac{Q_j}{Q_i} \right] \\ &= -k_B T \ln \left[ \frac{\int_{\Gamma_j} e^{-\beta H_j(\vec{p},\vec{q})} \, \mathrm{d}\vec{p}\mathrm{d}\vec{q}}{\int_{\Gamma_i} e^{-\beta H_i(\vec{p},\vec{q})} \, \mathrm{d}\vec{p}\mathrm{d}\vec{q}} \right]. \end{aligned} \tag{3.10}$$

[2] Note the temperature was not mentioned as a possible variable for transformation, it will later be explained why.

In the previous expression, the Hamiltonians and accessible phase spaces are state-dependent. A common approximation for many free energy methods is the assumption the two accessible phase spaces are identical, which can cause difficulties for complicated transformations. Eq. 3.10 will serve as the central expression for deriving the various methods upon which free energy calculations rely, due to its direct relationship to the microscopic ensembles, which can be generated by molecular simulations. Finally, it is now worth noting that a transformation in temperature, i.e. heating up or cooling down the system, is not possible as the free energy calculation would no longer be a ratio between partition functions as shown in Eq. 3.10, but instead the difference between the logarithmic partition functions, and thus face the same issues discussed for determining free energies using Eq. 3.9.

In the upcoming, we will look at how various methods and strategies can be adapted from Eq. 3.10 to determine the free energy between systems. In particular we will discuss the three major families of free energy methods namely *direct sampling methods*, *integration methods* and *perturbation methods*. It is worth mentioning two other large families of free energy methods that exist, namely *biased-equilibrium methods* and *non-equilibrium methods*, however, these will not be treated in detail here.

### 3.3.1 The Direct Sampling Method

The secret of the direct sampling method is almost given by the very name: By simply counting the number of occurrences of a binary criteria the free energy can be estimated. This scheme can be derived by multiplying Eq. 3.10 with the union of the two partition function's phase space ($\Gamma_{ij} = \Gamma_i \cup \Gamma_j$)

$$\Delta F_{i \to j}(N, V, T) = -k_B T \ln \left( \frac{Q_j Q_{ij}}{Q_i Q_{ij}} \right) = -k_B T \ln \left( \frac{P_j}{P_i} \right), \tag{3.11}$$

where $Q_{ij}$ is the united partition function of the partition functions $Q_i$ and $Q_j$. Since the probability is given by the number of samples of a specific event over the total number of samples, the total number of samples cancels in the expression. This method is however only useful given that the two states $i$ and $j$ are sampled with a sufficient transition frequency to obtain reliable statistics. This is usually the case when both the thermodynamic barrier, i.e. the free energy difference between the states, and the kinetic barrier, i.e. the free energy of activation, is fairly low. The direct sampling method is most commonly applied with the free energy of structural properties or simple binding equilibria. To differentiate the states into the binary count one commonly utilize a reduced measure, such as distances between specific atoms, angles, root-mean-square deviations (RMSD), cluster size, or even reduced variables obtained from dimensionality reduction techniques such as principal component analysis (PCA) or time-lagged independent component analysis (TICA). This has been applied to determine the pH-dependent rotameric and distance-

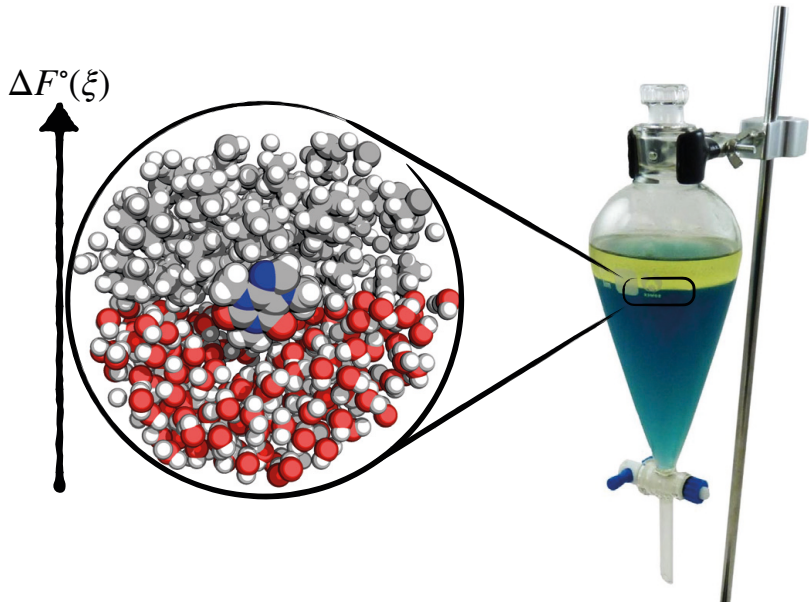



**Figure 3.3:** Illustration of the partitioning of caffeine between the aqueous phase and organic phase. In this partitioning experiment a separation funnel with colored water, occupying the bottom half, and colored cyclohexane, occupying the upper half, with caffeine occupying the two phases with a specific equilibrium. The standard Gibbs free energy for the partition of caffeine has experimentally been found to 12.9 kJ/mol [108] using the relationship $\Delta G^{\circ}_{\mathrm{trs}} = -RT\ln(P_i/P_j) = -RT\ln\left(\frac{[\mathrm{caffine}]_{\mathrm{cyclohexane}}}{[\mathrm{caffine}]_{\mathrm{water}}}\right)$. The insert visualizes a possible configuration obtainable from molecular simulations in which caffeine is found at the water-cyclohexane interface.

dependent structural equilibrium between two titratable residues utilizing first side chain dihedral angles and atom-atom distances found in paper I [46] and to quantify the structural contribution to the free energy of binding of $K^+$ to bis(crown ether) as found in paper III.

The direct sampling free energy method can be generalized to include non-binary states within the ensemble. Taking the most populated state as dictated by some collective variable $\xi(\vec{r}) = \xi(\vec{r_1}, \vec{r_2}, ..., \vec{r_N})$, as the ground state, Eq. 3.11 can be written as

$$\Delta F(\xi) = -k_B T \ln P(\xi). \tag{3.12}$$

Eq. 3.12 is commonly known as the *potential of mean force* and allows the free energy calculations along any reaction coordinate defined by the collective variable $\xi$, with the free energy being an ever-positive function, due to the definition of the ground state being the most populated state.

To illustrate the difficulty of the application of the direct sampling methods to obtain free energies for solvation thermodynamics, consider a partitioning experiment in which we wish to desire to know the equilibrium distribution of caffeine in an organic phase of cyclohexane and the aqueous phase of water, as visualized in Fig. 3.3. A naive approach to

calculating the free energy would be the creation of a simulation in which the upper half of the box is filled with cyclohexane and the bottom half of the box is filled with water, and allow periodic boundary conditions at the sides of the box, representing the liquid interface. With some random initial configuration of caffeine placed in either of the two phases, we can generate configurations using molecular simulations of caffeine's diffusion in the system, and thus simply calculate the free energy by counting the duration caffeine was in the cyclohexane phase and the water phase using Eq. 3.11, as it can be distinguished by the mid-plane of the simulation box. Furthermore one can define the reaction coordinate $\xi$ as the positive and negative displacement away from the interface and by histogramming determine the potential of mean force using Eq. 3.12. One major issue with this methodology is the slow diffusion of caffeine across the interface and generally infrequent transition between the two phases, rendering the need for unrealistic simulation times. Consequently one commonly utilizes either non-equilibrium methods such as steered dynamics, in which the caffeine molecule is pulled from one end of the system to another through the interface by an external force, or by biased-equilibrium sampling in which it is continuously made energetically unfavorable to occupy the same configurational ensemble of states.

### 3.3.2 Thermodynamic Integration & The Kirkwood Charging Formula

The thermodynamic integration method utilizes, that the free energy between two states of a system can be written as an integral of the work required to go the initial state to the final state as long as the change is done reversibly. Due to the free energy being a state function, the resulting free energy difference is independent of the path chosen from the initial to the final state and can be physical or non-physical. The generalized thermodynamic integration identity can be derived by the construction of a state-dependent Hamiltonian, $H\left(\vec{p},\vec{q},\lambda\right)$ where $\lambda$ is an arbitrary coupling-parameter linking the states of the system. Since the Hamiltonian is a function of $\lambda$, the free energy and partition function are also functions of $\lambda$ and we can therefore write the derivative of the free energy with respect to $\lambda$

$$\begin{aligned}\frac{\mathrm{d}F}{\mathrm{d}\lambda} &= -\left[\beta Q(N,V,T,\lambda)\right]^{-1}\frac{\mathrm{d}Q(N,V,T,\lambda)}{\mathrm{d}\lambda}\\ &= \frac{\int_\Gamma\left[\partial H\left(\vec{p},\vec{q},\lambda\right)/\partial\lambda\right]e^{-\beta H(\vec{p},\vec{q})}\,\mathrm{d}\vec{p}\mathrm{d}\vec{q}}{\int_\Gamma e^{-\beta H(\vec{p},\vec{q})}\,\mathrm{d}\vec{p}\mathrm{d}\vec{q}}\\ &= \left\langle\frac{\partial H\left(\vec{p},\vec{q},\lambda\right)}{\partial\lambda}\right\rangle_\lambda.\end{aligned} \tag{3.13}$$

In Eq. 3.13 we can see the free energy can be expressed as an ensemble average for a system with Hamiltonian $H(\lambda)$ which is highly appealing, due to it being a direct observable from molecular simulations. Taking the initial state to be $\lambda = 0$ and the end state to be $\lambda = 1$,

the difference in free energy is given by integrating Eq. 3.13

$$\Delta F = F(\lambda = 1) - F(\lambda = 0) = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda \, \mathrm{d}\lambda. \tag{3.14}$$

Eq. 3.14 is the generalized expression for thermodynamic integration. The integration is practically conducted by obtaining a set of ensemble averages from a series of simulations of different $\lambda$-values, with the choice of $\lambda$-values being completely arbitrary, as long as the desired end states can be reached. It was previously stated that the Hamiltonian was supposed to be a function of $\lambda$, while the only requirements to the function are it is differentiable and satisfied the boundary conditions of being state independent at $\lambda = 0$ and $\lambda = 1$, the Hamiltonian is most commonly taken to be a linear function on the form

$$H(\lambda) = H_i + \lambda \left( H_j - H_i \right). \tag{3.15}$$

Here the indices $i$ and $j$ refers to the thermodynamic state of the system before and after the transformation respectively. The usage of a linear combination of the two Hamiltonian has mainly two advantages. The first advantage is the calculation of the ensemble average derivative becomes a trivial task. Inserting Eq. 3.15 into Eq. 3.14 one obtains

$$\Delta F = \int_0^1 \langle H_j - H_i \rangle_\lambda \, \mathrm{d}\lambda, \tag{3.16}$$

which can be rapidly obtained as the converged averaged energy from a simulation simulated at any given $\lambda$-value. The second advantage is the Gibbs-Bogoliubov inequality can be shown to apply for the linear coupling scheme, thus dictating that the ensemble average derivative can never increase with increasing $\lambda$ and can thus be utilized to test the validity of the simulation results.[40]

To calculate the solvation free energy, and thus chemical potential, for a vacuum-to-solution solvation process we can utilize thermodynamic integration to slowly grow in a solute molecule using a series of $\lambda$-values, with the coupling-parameter coupled to the interactions between the solute and the surrounding solvent. Taking the Hamiltonian to be a linear function of $\lambda$, as shown in Eq. 3.15, one could write the solvent-solute Hamiltonian as

$$H(\lambda) = \lambda \sum_i^N \sum_{j=i+1}^N 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}}, \tag{3.17}$$

where the first term is the well-known Lennard-Jones potential and the second term a simplified Coulomb potential. The linearly scaling however is often ill-behaved at $\lambda$ values near the end states in which the solute particle is in one step going from being completely absent from the system to now exclude volume creation a singularity (infinite potential) when the solvent is overlapping with the solute, thus causing large fluctuations in $\partial H(\lambda)/\partial \lambda$.[11,123]

Consequently it is today common practice to utilize potentials which smoothens out the singularity using so-called "softcore potentials".[11,126,150] Similarly it is also important to constantly maintain an exclude volume for particles possessing charge, as oppositely charged particles would otherwise experience an electrostatic singularity. Besides the issues associated with the potentials, systematic errors are inevitable, due to the discretization of the integral, in particular, if $\partial H(\lambda)/\partial\lambda$ varies greatly in some ranges of $\lambda$ it must be sampled extra rigorously.

The thermodynamic integration identity (Eq. 3.14) was first discovered by John G. Kirkwood, however with a different goal in mind; to express the chemical potential of the components in a liquid solution in terms of molecular pair distribution functions.[56] This culminated in the well known *Kirkwood charging formula*. To understand the formula, consider a species-decomposed $\lambda$-coupled potential energy function

$$U(\lambda) = u_s + \sum_{i=1}^{N_{sw}} u_{sw}(\lambda, i) + u_{ww}, \tag{3.18}$$

where $u_s$ is the potential energy of the solute being a one-body term, $u_{sw}(\lambda, i)$ is the potential energy arising from solute-solvent interactions, and $u_{ww}$ is the potential energy arising from the solvent-solvent interactions, with the last two terms being pair potentials. By this definition we can via our coupling-parameter $\lambda$ introduce interactions between the solute and solvent, as it was characteristic for a solvation process. Inserting Eq. 3.18 into Eq. 3.13 one obtains

$$\begin{aligned}\frac{\partial\mu^{\text{ex}}}{\partial\lambda} &= \left\langle \sum_{i=1}^{N_w} \frac{\partial u_{sw}(\lambda, i)}{\partial\lambda} \right\rangle_{\lambda} \\ &= \int_{-\infty}^{\infty} \frac{\partial u_{sw}(\lambda, i)}{\partial\lambda} \left\langle \sum_{i=1}^{N_w} \delta(\vec{r}_w - \vec{r}_{w,i})\delta(\vec{r}_s - \vec{r}_{s,i}) \right\rangle \mathrm{d}\vec{r}_s \, \mathrm{d}\vec{r}_w.\end{aligned} \tag{3.19}$$

Here we have utilized only the solute-solvent potential energy $u_{sw}$ is varying with the coupling-parameter $\lambda$, and changed the order of integration such that the ensemble average is taken over the summation of $\delta$-functions. The ensemble average can be identified as the equilibrium solute-solvent density distribution function $\rho_{sw}$.[45] Substituting in the density distribution function and integrating over $\lambda$ we obtain the Kirkwood charging formula

$$\Delta\mu = \int_0^1 \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\partial u_{sw}(\lambda, i)}{\partial\lambda} \rho_{sw}(\vec{r}_s, \vec{r}_w) \, \mathrm{d}\vec{r}_s \, \mathrm{d}\vec{r}_w. \tag{3.20}$$

While Eq. 3.20 is exact, it can not be applied easily due to the solute-solvent density distribution function is represented over a high-dimensional set of coordinates of positions.

Instead one can utilize a projection coordinate such as the radial distribution function utilized in the methods dubbed reference interaction site model (RISM), or energy distribution functions utilized in the method of energy-representation theory of solvation (covered in section 3.3.4), such that the high-dimensional expression can be omitted to arrive at useful expressions.

### 3.3.3 Free Energy Perturbation & The Widom Particle-Insertion Method

In Eq. 3.13 we saw the free energy could be expressed as a derivative with respect to the coupling-parameter $\lambda$ and be integrated to yield the difference in free energy. An alternative approach would be to express the derivative as a difference between $\lambda$ and a finite increment in the coupling-parameter $\lambda + \Delta\lambda$

$$\frac{\mathrm{d}F(\lambda)}{\mathrm{d}\lambda} = \frac{F(\lambda + \Delta\lambda) - F(\lambda)}{\Delta\lambda}. \tag{3.21}$$

Utilizing the expression for the absolute free energy (Eq. 3.9) we can rewrite Eq. 3.21 to

$$\begin{aligned}
\frac{\mathrm{d}F(\lambda)}{\mathrm{d}\lambda} &= -k_B T \frac{\ln\left[\int_\Gamma e^{-\beta H(\vec{p},\vec{q},\lambda+\Delta\lambda)}\,\mathrm{d}\vec{p}\mathrm{d}\vec{q}\right] - \ln\left[\int_\Gamma e^{-\beta H(\vec{p},\vec{q},\lambda)}\,\mathrm{d}\vec{p}\mathrm{d}\vec{q}\right]}{\Delta\lambda} \\
&= \frac{-k_B T}{\Delta\lambda} \ln\left[\frac{\int_\Gamma e^{-\beta H(\vec{p},\vec{q},\lambda+\Delta\lambda) - H(\vec{p},\vec{q},\lambda)} e^{-\beta H(\vec{p},\vec{q},\lambda)}}{\int_\Gamma e^{-\beta H(\vec{p},\vec{q},\lambda)}}\right] \\
&= \frac{-k_B T}{\Delta\lambda} \ln\left\langle e^{-\beta H(\vec{p},\vec{q},\lambda+\Delta\lambda) - H(\vec{p},\vec{q},\lambda)}\right\rangle_\lambda .
\end{aligned} \tag{3.22}$$

In the second equality, the expression was added with ratio $\ln\left(\frac{Q(\lambda)}{Q(\lambda)}\right)$, followed by the factoring of the unperturbed Boltzmann factor. We here see the reasoning behind the names "free energy perturbation" and "exponential averaging" (being another common name), as the free energy can be found, from the exponential energy difference between the perturbed and unperturbed system, averaged over the unperturbed phase space, i.e. an averaging of Boltzmann factors. The free energy between the individual coupling-parameter increments can thus be found as the sum of the individual increments between the end states

$$\Delta F = -k_B T \sum_{\lambda=0}^{\lambda=1} \ln\left\langle e^{-\beta H(\vec{p},\vec{q},\lambda+\Delta\lambda) - H(\vec{p},\vec{q},\lambda)}\right\rangle_\lambda . \tag{3.23}$$

Within the framework of Eq. 3.23 one may arbitrarily choose the number and size of the increments between the final states, with the possibility to even do a full perturbation between the final states in a single increment, in which the summation then disappears from the expression.

Within the topic of solvation thermodynamics and the determination of chemical potentials, the usage of single increment free energy perturbation is best manifested in the Widom particle insertion method,[145] in which the coupling-parameter is taken to the binary option of having no solute particles $\lambda = 0$, and one solute particle $\lambda = 1$ with the given solvent remaining fixed for both conditions. This approach can be utilized on the fly during molecular simulation, such that an unperturbed simulation is periodically paused with the insertion of a trial particle in a random position, and average the Boltzmann factor arising from the difference in energy between the perturbed and unperturbed state.[40,145] The Widom particle insertion method is a very simple, yet useful scheme for the computation of chemical potentials and has thus undergone continuous development since the initial publication of the method by Widom, to also include the possibility of obtaining chemical potentials by a particle deletion, i.e. the reverse process of the Widom method, and by particle reinsertion, i.e. to place the particles ones more into the system after a deletion, providing various methods to improve the sampling.[12] However despite its simplicity, single increment free energy perturbations methods, are known to perform poorly, in particular, if the perturbation is large, such as the insertion of a particle into a dense fluid, creating configurations with particle overlap characterized by high energy, thus causing the explored phase spaces to not overlap, which could be mediated by a staged insertion using multiple increments. This problem is in particular related to the low probability (a rare event) of spontaneously creating a cavity suitable for the solute molecule, with the larger solutes requiring a larger cavity. Another complication is that of molecules that are characterized by multiple internal degrees of freedom, which may be highly correlated with the solvent composition and coordination, as it is for example seen in proteins, are decoupled.

With free energy perturbation sampling the unperturbed state and estimating energy difference between the perturbed and unperturbed state, systematic errors are expected to occur as the forward variation (from 0 to 1) might not yield the negative backward variation (from 1 to 0) as otherwise dictated due to the free energy being a state function. Therefore modern free energy perturbation commonly utilizes the averaging of forwarding and backward calculations as previously mentioned for the Widom methodology. However, it has been found that itself can also be a source of systematic error.[68] Consequently modern studies are more commonly applying *Bennett's acceptance ratio methods*,[8,121] which are characterized by utilizing both the forward and backward variation of $\lambda$-values, while simultaneously minimizing the standard error for a given simulation time of the specific $\lambda$-values using weighting-functions.[8]

### 3.3.4 Density Functional Theory & Energy-representation Theory of Solvation

In this section, we will present the theory for the computation of the excess chemical potential using the energy-representation theory of solvation developed by the Matubayasi

group in Osaka, Japan. The method is characterized in defining a new reaction coordinate, in specific the solute-solvent pair-energy, which combined with the Kirkwood charging formula and classical density functional theory yields an expression for the calculation of excess chemical potentials relying only on end-state molecular simulations.

**Definitions Within the Energy-representation Theory**

**1. The Energy Collective Coordinate**

Within the theory of energy-representation, we introduce a new collective variable/coordinate namely the pair interaction energy between a solute molecule and a solvent molecule donated $\epsilon$. This coordinate is taken to be $\lambda$-independent and thus must be calculated with the solute-solvent potential at full coupling corresponding to taking $\lambda = 1$ (c.f. Eq. 3.18). Consequently, the pair solute-solvent energy is defined as:

$$v_{sw}(\vec{r}_s, \vec{r}_w) = u_{sw}(\lambda = 1, \vec{r}_s, \vec{r}_w). \tag{3.24}$$

**2. The Microscopic Energy Density**

The instantaneous, i.e. a single arbitrary particle configuration, pair-energy distribution function is introduced as

$$\begin{aligned} \hat{\rho}^{\epsilon}_{sw}(\epsilon) &= \sum_{i=0}^{N_w} \delta(v_{sw}(\vec{r}_s, \vec{r}_w) - \epsilon) \\ &= \int_{-\infty}^{\infty} \delta(v_{sw}(\vec{r}_s, \vec{r}_w) - \epsilon)\hat{\rho}_{sw}(\vec{r}_s, \vec{r}_w)\, \mathrm{d}\vec{r}_s\, \mathrm{d}\vec{r}_w. \end{aligned} \tag{3.25}$$

**3. The Energy-representation's Energy Equation**

Just like the mean excess energy can be expressed given the knowledge on the pair-energies of the system and the radial distribution function as it is found from the "energy equation"[45] the potential energy within the energy-representation can be similarly written as

$$u^{\epsilon}_{sw}(\epsilon) = \int_{-\infty}^{\infty} \delta(v_{sw}(\vec{r}_s, \vec{r}_w) - \epsilon)u_{sw}(\lambda, \vec{r}_s, \vec{r}_w)\, \mathrm{d}\vec{r}_s\, \mathrm{d}\vec{r}_w. \tag{3.26}$$

We choose the $\lambda$-path to be one in which $u^{\epsilon}_{sw}(\lambda, \vec{r}_s, \vec{r}_w)$ is an equi-energy surface of $v_{sw}(\vec{r}_s, \vec{r}_w)$ by the restraint: $u^{\epsilon}_{sw}(\lambda, \vec{r}_s, \vec{r}_w) = \lambda v_{sw}(\vec{r}_s, \vec{r}_w)$, thus allowing us to express the $\lambda$-coupled pair energy as

$$u_{sw}(\lambda, \vec{r}_s, \vec{r}_w) = \int_{-\infty}^{\infty} \delta(v_{sw}(\vec{r}_s, \vec{r}_w) - \epsilon)u^{\epsilon}_{sw}(\epsilon)\, \mathrm{d}\epsilon. \tag{3.27}$$

**4. The Solute-solvent Ensemble Density Distribution Function In the Energy Representation**

The solute-solvent ensemble density distribution function in the canonical ensemble ($NVT$) ensemble is given by:

$$\rho^{\epsilon}_{sw}(\lambda,\epsilon) = \langle\hat{\rho}(\epsilon)\rangle_{\lambda}. \tag{3.28}$$

Using the definition of the microscopic energy density and rewriting the ensemble averages, we find

$$\begin{aligned}\rho^{\epsilon}_{sw}(\lambda,\epsilon) =& \frac{\int_{\Gamma}\left[\int_{-\infty}^{\infty}\delta(v_{sw}(\vec{r}_s,\vec{r}_w)-\epsilon)\hat{\rho}_{sw}(\vec{r}_s,\vec{r}_w)\,\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w\right]}{\int_{\Gamma}e^{-\beta U(\lambda,\vec{r}_s,\vec{r}_w)}\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w^{N_w}}\times \\ & e^{-\beta U(\lambda,\vec{r}_s,\vec{r}_w)}\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w^{N_w},\end{aligned} \tag{3.29}$$

where the first integral over $\Gamma$ is the integration over phase space. Conducting a change in the integration order we can rewrite the equation as

$$\begin{aligned}\rho^{\epsilon}_{sw}(\lambda,\epsilon) = \int_{-\infty}^{\infty}\delta(v_{sw}(\vec{r}_s,\vec{r}_w)-\epsilon)\,\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w\times \\ \frac{\int_{\Gamma}\left[\hat{\rho}_{sw}(\vec{r}_s,\vec{r}_w)\right]e^{-\beta U(\lambda,\vec{r}_s,\vec{r}_w)}\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w^{N_w}}{\int_{\Gamma}e^{-\beta U(\lambda,\vec{r}_s,\vec{r}_w)}\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w^{N_w}}.\end{aligned} \tag{3.30}$$

The fraction can be identified as the Boltzmann ensemble average of the instantaneous solute-solvent density distribution, and is thus equal to the the solute-solvent density distribution $\rho_{sw}(\vec{r}_s,\vec{r}_w)$:

$$\rho^{\epsilon}_{sw}(\lambda,\epsilon) = \int_{-\infty}^{\infty}\delta(v_{sw}(\vec{r}_s,\vec{r}_w)-\epsilon)\rho_{sw}(\vec{r}_s,\vec{r}_w)\,\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w. \tag{3.31}$$

The starting point to solve the Kirkwood charging formula (Eq. 3.20) will be the rewriting of the function to depend on the energy collective coordinate and its associated distribution function. To do so we substitute the partial derivative of Eq. 3.27 and substitute it into the Kirkwood charging formula:

$$\Delta\mu = \int_0^1\mathrm{d}\lambda\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\delta(v_{sw}(\vec{r}_s,\vec{r}_w)-\epsilon)\frac{\partial u^{\epsilon}_{sw}(\epsilon)}{\partial\lambda}\,\mathrm{d}\epsilon\right]\rho_{sw}(\vec{r}_s,\vec{r}_w)\,\mathrm{d}\vec{r}_s\,\mathrm{d}\vec{r}_w. \tag{3.32}$$

By changing the order of integration and substituting the solute-solvent density distribution

function in the energy representation (Def. 4), we obtain

$$\begin{aligned}\Delta\mu &= \int_0^1 \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\partial u_{sw}^{\epsilon}(\epsilon)}{\partial\lambda}\, \mathrm{d}\epsilon \left[\int_{-\infty}^{\infty} \delta(v_{sw}(\vec{r}_s,\vec{r}_w)-\epsilon)\rho_{sw}(\vec{r}_s,\vec{r}_w)\, \mathrm{d}\vec{r}_s\, \mathrm{d}\vec{r}_w\right] \\ &= \int_0^1 \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\partial u_{sw}^{\epsilon}(\epsilon)}{\partial\lambda}\rho_{sw}^{\epsilon}(\lambda,\epsilon)\, \mathrm{d}\epsilon.\end{aligned} \tag{3.33}$$

Eq. 3.33 is the Kirkwood charging formula within the energy representation formulation, due to the only dependency of the potential and the ensemble density distribution function on the solute-solvent pair energy coordinate. The Kirkwood charging formula can be further solved

$$\begin{aligned}\Delta\mu &= \int_{-\infty}^{\infty} \mathrm{d}\epsilon \int_0^1 \frac{\partial u_{sw}^{\epsilon}(\epsilon)}{\partial\lambda}\rho_{sw}^{\epsilon}(\lambda,\epsilon)\, \mathrm{d}\lambda \\ &= \int_{-\infty}^{\infty} \mathrm{d}\epsilon \left[u_{sw}^{\epsilon}(\lambda=1,\epsilon)\rho_{sw}^{\epsilon}(\lambda=1,\epsilon) - \int_0^1 \frac{\partial\rho_{sw}^{\epsilon}(\lambda,\epsilon)}{\partial\lambda}u_{sw}^{\epsilon}(\epsilon)\, \mathrm{d}\lambda\right] \\ &= \int_{-\infty}^{\infty} \rho_{sw}^{\epsilon}(\lambda=1,\epsilon)\epsilon\, \mathrm{d}\epsilon - \int_0^1 \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\partial\rho_{sw}^{\epsilon}(\lambda,\epsilon)}{\partial\lambda}u_{sw}^{\epsilon}(\epsilon)\, \mathrm{d}\epsilon,\end{aligned} \tag{3.34}$$

where we have in the first equality changed the order of integration, in the second equality integrated by parts for the inner integral, and in the third equality reverted the order of integration while rewriting the expression using definition 1 and 3 i.e. $u_{sw}^{\epsilon}(\lambda=1,\epsilon) = v_{sw}^{\epsilon} = \epsilon$. We will now donate the second term as a functional of the potential and the solute-solvent ensemble density distribution function:

$$\mathcal{F}\left[\rho_{sw}^{\epsilon}(\lambda,\epsilon), u_{sw}^{\epsilon}(\epsilon)\right] = \int_0^1 \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\partial\rho_{sw}^{\epsilon}(\lambda,\epsilon)}{\partial\lambda}u_{sw}^{\epsilon}(\epsilon)\, \mathrm{d}\epsilon. \tag{3.35}$$

To find an expression for the functional of the potential and the solute-solvent ensemble density distribution function we continue by decomposing the ensemble density distribution function into a *direct* contribution arising from particle pair interaction, and an *indirect* contribution, arising from the correlation of particles intermediate to the pair.[45]

$$\rho_{sw}^{\epsilon}(\lambda,\epsilon) = \rho_{sw}^{\epsilon}(\lambda=0,\epsilon)e^{-\beta(u_{sw}^{\epsilon}(\epsilon)+w_{sw}^{\epsilon}(\epsilon))}, \tag{3.36}$$

and similarly we can rewrite the potential to depend on the indirect contribution

$$u_{sw}^{\epsilon}(\epsilon) = -k_B T \ln\left(\frac{\rho_{sw}^{\epsilon}(\lambda,\epsilon)}{\rho_{sw}^{\epsilon}(\lambda=0,\epsilon)}\right) - w_{sw}^{\epsilon}(\epsilon). \tag{3.37}$$

Using these definitions we can evaluate the functional expression to solve the Kirkwood charging formula within the energy representation. Inserting the above expression into Eq.

3.35 we can write the functional as

$$\mathcal{F}\left[\rho^{\epsilon}_{sw}(\lambda,\epsilon),u^{\epsilon}_{sw}(\epsilon)\right]=\int_{-\infty}^{\infty}\mathrm{d}\epsilon\int_{0}^{1}\frac{\partial\rho^{\epsilon}_{sw}(\lambda,\epsilon)}{\partial\lambda}\times \left(-k_BT\ln\left(\frac{\rho^{\epsilon}_{sw}(\lambda,\epsilon)}{\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)}\right)-w^{\epsilon}_{sw}(\epsilon)\right)\mathrm{d}\lambda, \tag{3.38}$$

where we have changed the order of integration. The first integral can be solved analytically and the functional can thus be found to be equal to

$$\begin{aligned}\mathcal{F}[\rho^{\epsilon}_{sw}(\lambda,\epsilon),u^{\epsilon}_{sw}(\epsilon)]=&k_BT\int_{-\infty}^{\infty}\Bigg[\big(\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)-\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)\big)-\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)\times\\ &\ln\frac{\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)}{\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)}-\beta\int_{0}^{1}\frac{\partial\rho^{\epsilon}_{sw}(\lambda,\epsilon)}{\partial\lambda}w^{\epsilon}_{sw}(\lambda,\epsilon)\mathrm{d}\lambda\Bigg]\mathrm{d}\epsilon.\end{aligned} \tag{3.39}$$

The above expression may be further simplified if we choose the $\lambda$-dependency of the potential such that the ensemble density distribution is a linear combination of $\lambda$ end-states:

$$\rho^{\epsilon}_{sw}(\lambda,\epsilon)=\lambda\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)+m(1-\lambda)\rho^{\epsilon}_{sw}(\lambda=0,\epsilon) \tag{3.40}$$

thus allowing the partial derivative to be rewritten as

$$\begin{aligned}\mathcal{F}[\rho^{\epsilon}_{sw}(\lambda,\epsilon),u^{\epsilon}_{sw}(\epsilon)]=&k_BT\int_{-\infty}^{\infty}\Bigg[\big(\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)-\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)\big)-\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)\times\\ &\ln\frac{\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)}{\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)}-\beta\big(\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)-\rho^{\epsilon}_{sw}(\lambda=0,\epsilon)\big)\times\\ &\int_{0}^{1}w^{\epsilon}_{sw}(\lambda,\epsilon)\mathrm{d}\lambda\Bigg]\mathrm{d}\epsilon.\end{aligned} \tag{3.41}$$

With the chemical potential being ultimately given by

$$\Delta\mu=\int_{-\infty}^{\infty}\rho^{\epsilon}_{sw}(\lambda=1,\epsilon)\epsilon\,\mathrm{d}\epsilon-\mathcal{F}[\rho^{\epsilon}_{sw}(\lambda,\epsilon),u^{\epsilon}_{sw}(\epsilon)]. \tag{3.42}$$

Eq. 3.42 is the fundamental equation for the determination of the chemical potentials within the energetic representation theory and it should be noted Eq. 3.42 is an exact expression with no approximations having been made up to this point. Physically Eq. 3.42 also states the chemical potential can be thought of as two contributions: One being the contribution arising from the interactions between the solute and solvent as the first term can be identified as the mean pair-energy between the solute and solvent at full coupling ($\lambda=1$), while the second contribution is associated with the energetics of cavity formation and in turn solvent reorganization due to the functional depending on the pair-energy

ensemble density distribution for the system being fully uncoupled and fully coupled. The practical limitation of Eq. 3.42 arises from the $\lambda$-dependency of the indirect part of the potential of mean force within the functional. To eliminate this dependency, approximate functionals can be applied.

Using Percus's method of functional expansion,[97] it can be found that the indirect potential of mean force $\lambda$-integral can be solved analytically using the Percus-Yevick (PY) like and hypernetted-chain (HNC) like approximations.[41] However it has been recognized in the case of simple liquids that the PY approximation performs better in the case of short-range repulsive potentials, while the HNC approximation performs better in the case of long-range attractive potentials.[45] Consequently Matubayasi and Nakahara[77] choose to construct a hybrid functional as a combination of the two approximations, in which the PY-like approximation is utilized in the unfavorable energy-region of solvation ($w_{sw}^{\epsilon} \geq 0$) and the HNC-like approximation is utilized in the favorable energy-region of solvation ($w_{sw}^{\epsilon} < 0$). However the PY-like and HNC-like approximation are related to one another, and henceforth we shall only discuss the hybrid functional in terms of the HNC-like approximation. An alternative to the determination of the indirect potential of mean force by approximate functionals is the direct sampling by molecular simulations like molecular dynamics and Monte Carlo simulations for interactions between solute and solvent not characterized by Pauli repulsion i.e. outside the solute-core region. Matubayasi consequently proposed to use the HNC-like approximation only when it can not be determined from molecular simulations.[112] For the core-region of energies, which are unsampled by the usage of molecular simulations at $\lambda = 1$, the probability density distribution must be much greater at zero coupling ($\rho_{sw}^{\epsilon}(\lambda = 1) \ll \rho_{sw}^{\epsilon}(\lambda = 0)$), and hence the dependency on $\rho_{sw}^{\epsilon}(\lambda = 1)$ disappear in the HNC-like approximation for the solute-core region. Instead, the core region is calculated in the ensemble where the solute and solvent are fully decoupled ($\lambda = 0$), with the simple approach being the insertion of solute molecules in random orientations, into the ensemble of pure solvent configurations. Skipping the full derivation of the PY-like and HNC-like functionals and their individual $\lambda$-integration, the $\lambda$-integration over the indirect potential of mean force in the method by Matubayasi, Nakahara, and Sakuraba is given by[112]

$$\beta \int_0^1 w_{sw}^{\epsilon}(\lambda, \epsilon)\, \mathrm{d}\lambda \approx \alpha(\epsilon) F_w + [1 - \alpha(\epsilon)] F_{w,\mathrm{HNC}}, \tag{3.43}$$

where the functions $F_w$ and $F_{w,\mathrm{HNC}}$ are written as functions of the PY-like and NHC-like expressions for the $\lambda$-integral

$$F_w = \begin{cases} \frac{\beta w_{sw}^{\epsilon}(\epsilon)}{2}, & \text{when } w_{sw}^{\epsilon}(\epsilon) \geq 0 \\ \beta w_{sw}^{\epsilon}(\epsilon) + 1 + \frac{\beta w_{sw}^{\epsilon}(\epsilon)}{e^{-\beta w_{sw}^{\epsilon}(\epsilon)} - 1}, & \text{when } w_{sw}^{\epsilon}(\epsilon) < 0, \end{cases} \tag{3.44}$$

and

$$F_{w,\mathrm{HNC}} = \begin{cases} \frac{\beta w_{sw}^{\epsilon,\mathrm{HNC}}(\epsilon)}{2}, & \text{when } w_{sw}^{\epsilon}(\epsilon) \geq 0 \\ -\ln\left[1-\beta w_{sw}^{\epsilon,\mathrm{HNC}}(\epsilon)\right] + 1 + \frac{\ln\left[1-\beta w_{sw}^{\epsilon,\mathrm{HNC}}(\epsilon)\right]}{\beta w_{sw}^{\epsilon,\mathrm{HNC}}(\epsilon)}, & \text{when } w_{sw}^{\epsilon}(\epsilon) < 0, \end{cases} \tag{3.45}$$

and the parameter $\alpha(\epsilon)$ is responsible for merging the different indirect potential of mean force functions:

$$\alpha(\epsilon) = \begin{cases} 1, & \text{when } \rho_{sw}^{\epsilon}(\epsilon, \lambda = 1) \geq \rho_{sw}^{\epsilon}(\epsilon, \lambda = 0) \\ 1 - \left(\frac{\rho_{sw}^{\epsilon}(\epsilon,\lambda=1) - \rho_{sw}^{\epsilon}(\epsilon,\lambda=0)}{\rho_{sw}^{\epsilon}(\epsilon,\lambda=1) + \rho_{sw}^{\epsilon}(\epsilon,\lambda=0)}\right)^2, & \text{when } \rho_{sw}^{\epsilon}(\epsilon, \lambda = 1) < \rho_{sw}^{\epsilon}(\epsilon, \lambda = 0). \end{cases} \tag{3.46}$$

### 3.3.5 Thermodynamic Cycles

Due to the nature of thermodynamic state functions being path-independent for reversible processes, it opens the possibility to determine free energies of otherwise complicated processes of interest. Within protein chemistry, the majority of thermodynamic cycles usually rely on the usage of transfer free energies, in which for example a ligand and product are transferred from the solvent to the binding site of proteins to calculate the free energy of catalyzing a specific reaction in proteins. Another example could be the transfer of titratable amino acid analogs to the folded state of proteins to determine the p$K_a$ of titratable residues in proteins which is perturbed by the dehydration and specific interactions found in proteins as it is visualized in Fig. 3.4. Specific for the determination of p$K_a$ values in proteins we here see it is sufficient to know the p$K_a$ of the group in water yielding the free energy of (de)protonation in water and the difference in solvation free energy of the protonation state and dehydrated state to obtain the free energy of (de)protonation in the protein. Examples of thermodynamic cycles used in this work include the presented solvation/(de)protonation cycle previously presented and a diprotic acid cycle, in which an acid with two states can adapt two energetically different states in the pathway from the fully protonated state to the fully deprotonated state.

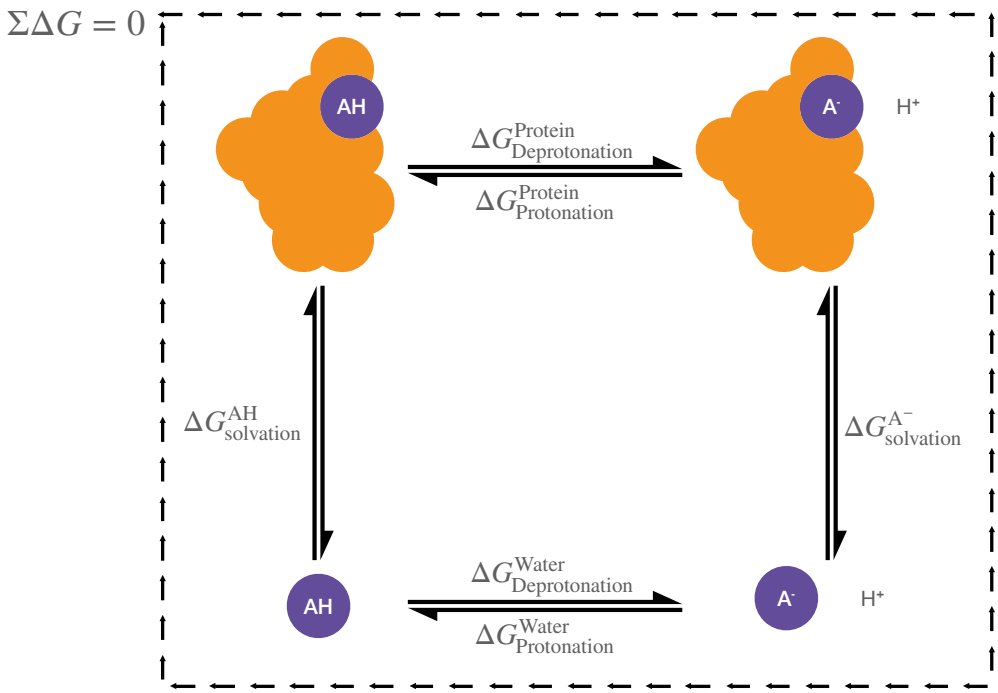



**Figure 3.4:** Thermodynamic cycle for the calculation of free energy of (de)protonation (upper horizontal axis), using solvation free energies (vertical axis) of the protonated (left) and deprotonated state (right). To complete the cycle knowledge on the free energy of (de)protonation in water is required which can be experimentally obtainable from example wise p$K_a$ measurements. The presented cycle is also generally applicable to studies of ligand binding as protons may be considered the smallest possible ligand.

# Chapter 4

# Molecular Simulations

*An intelligence which could, at any moment, comprehend all the forces by which nature is animated and the respective positions of the beings of which it is composed, and moreover, if this intelligence were far-reaching enough to subject these data to analysis, it would encompass in that same formula both the movements of the largest bodies in the universe and those of the lightest atom.*

— Pierre-Simon Laplace, 1749-1827

Simulations offer, like experiments, insight into the nature we are living in, however as with any experimental method, simulations also come with their advantage and their limitations. For example, within the category of experimental scattering techniques, we address "the resolution" of the experiment, which is the lowest distance at which we can distinguish structural features from one another. While the resolution in scattering experiments is an observable, in molecular simulations it is a choice made by the modeler. In particular, the levels of theory and coarse-graining of the molecular system on which the simulations are based reflect the resolution chosen (see Fig. 4.1). For example, upon going from a quantum mechanical description to a classical mechanical description we abandon the description of explicit electrons to instead utilize a mean-field description of electron-electron interactions through potential energy functions. This choice is logical if one is not interested in accurately describing electronic properties for larger molecules. The next level of coarse-graining usually either involves a mean-field description of a solvent surrounding a solute and/or the reduction of structural details in the solute through spheres composing multiple atoms or in the case of proteins; whole residues. As of consequence, we can design molecular simulations to match the experimental resolution, however with the cost of abandoning details below the chosen resolution. As stated in the introduction, statistical thermodynamics is required to provide the necessary interpretation of molecular phenomena to experimental observations, using molecular simulations we can inverse the process,

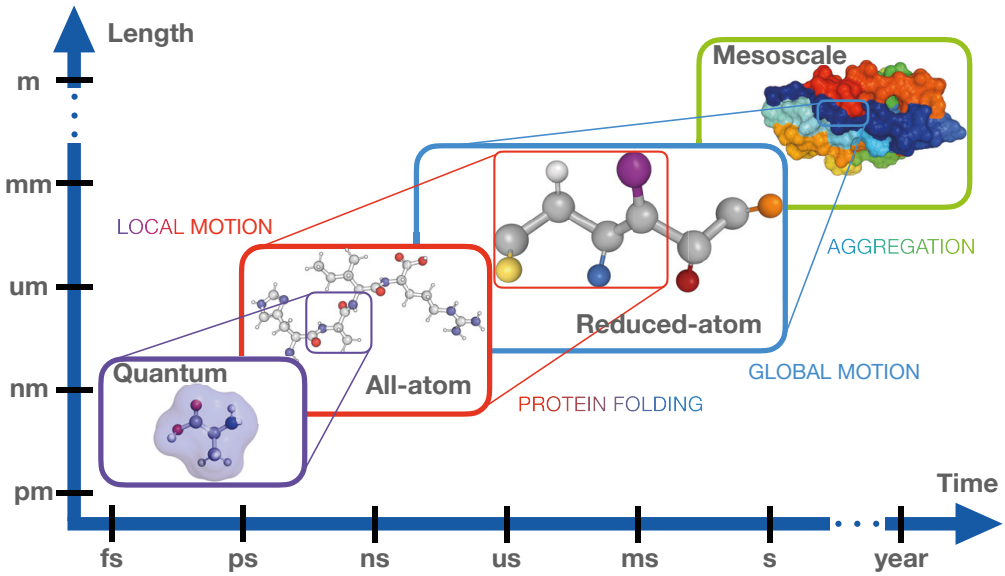



**Figure 4.1:** Illustrations of the various simulation methods and their characteristic usage in the time- and length-scale spectrum. At the highest level of resolution, we for example find quantum methods, which are limited to the studying properties occurring on the femtosecond to picosecond time scale and picometer to nanometer length scale. In contrast, we find all-atomistic molecular dynamics and Monte Carlo, where explicit electrons and their interactions have been replaced with effective pair-potentials to mimic electronic interactions, can simulate picosecond to microsecond time scale and on the nanometer to micrometer length-scale. Taking it a step further one may reduce the structural details by coarse-graining atomistic matter to simulate at even longer time scales and length scales. As such, the choice of resolution should reflect the properties of interest, like protein folding of fast-folding proteins and the local motion of proteins, which can be studied on the atomistic scale while larger globular motion and slow protein folding are best left for coarse-graining. Among the lowest processes are the formation of highly structured aggregates are studied on the mesoscale level. However, in the author's opinion, it is in the best of interest to push the limits of atomistic simulations to study larger and slower properties and fall back to coarse-graining when beyond our current capabilities.

and derive experimental observations using statistical thermodynamics. This bidirectional relationship between molecular simulations and experiments can assist to discover the underlying molecular mechanism of molecular processes. However, an inherent difficulty between molecular simulations and experiments, is the difference in size, time, and conditions at which the experiment was conducted. For example, if one were experimenting on a glass of water to find the heat capacity over the process of 1 minute, it would be equivalent to the study of approximately $10^{24}$ water molecules, using $10^{16}$ iterations by solving Newton's equations of motion, thus rendering it an impossible computational challenge. As of consequence, simulations are commonly conducted at a much smaller scale, thus creating elements of uncertainty, in which we rely on experimental verification to confirm the simulations can be utilized to extract information about the system and its properties.

In this chapter, we will focus on the main methods utilized in generating configurations with Boltzmann distributed probabilities, with the main methods being molecular dynamics including the stochastic Langevin dynamics and Markov chain Monte Carlo simula-

tions. Despite the methods being highly different in their approach to generating the statistical ensemble, the result should be the same, however as with everything; they each have their advantages and disadvantages. Finally, the combination of molecular dynamics and Monte Carlo simulations will be discussed, with the focus on the possibility to increase the efficiency of the methods and sampling of more exotic thermodynamic ensembles.

Before moving on, I feel the need to address the quote by Pierre-Simon Laplace, stated at the start of this chapter. In particular, the quote presented is from his work *Philosophical Essay on Probabilities* from 1814. In his work, Laplace was arguing for a deterministic image of the world based on classical mechanics, with this "intelligence" later having been named Laplace's demon, capable of knowing all particles positions and momenta and thus be able to predict future and past outcomes of systems. However, to my amusement, I like to think Laplace is very much describing a modern-day molecular dynamics simulation, with the "intelligence" being the modern computer.

## 4.1 Molecular Dynamics

The method Molecular Dynamics (MD) generates configurations by solving Newton's equations of motion thus propagating the movement of molecules and atoms over time. In particular, the dynamics are governed by Newton's second law of motion

$$\vec{F} = m\vec{a} = m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = m\frac{\mathrm{d}^2\vec{q}}{\mathrm{d}t^2}. \tag{4.1}$$

Where $\vec{F}$ is the force, $\vec{a}$ is the acceleration, $\vec{v}$ is the velocity, and $\vec{q}$ is the position with all the quantities being vectors, and $m$ being the mass of the object. While the first equality can be identified as Newton's second law of motion, the following equalities relate the kinematic variable acceleration to remaining kinematic variables namely the velocity and position, thus setting the framework for the kinematic equations, allowing the determination of the kinematic state at any given time. Except for very simplified, independent systems such as objects moving at constant acceleration, or moving according to simple energy functions like independent harmonic oscillators, Eq. 4.1 can not be solved analytically, and must instead be solved by numeric integration. The choice of numerical integration scheme to solve Eq. 4.1 is within molecular dynamics terminology called an *integrator* with the most common choice being the Verlet integrators which include the Störmer-Verlet method[138] also named position Verlet algorithm, and the velocity Verlet algorithm.[129] The velocity Verlet integration scheme is characteristic in utilizing a Taylor expansion around the time $t$ both forward and backward in time, thus effectively applying the midpoint method. The position and velocity at the time increment $t + \Delta t$ by velocity Verlet integration is given

by

$$\begin{aligned}\vec{q}(t+\Delta t) &= \vec{q}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)\Delta t^2 \\ \vec{v}(t+\Delta t) &= \vec{v}(t) + \frac{\vec{a}(t)+\vec{a}(t+\Delta t)}{2}\Delta t.\end{aligned} \tag{4.2}$$

The three most important features of the velocity Verlet integration scheme, making it highly appealing is I) it is self-starting, meaning that once the initial configuration (positions and velocities/momenta) of the system has been established, any future can be determined without the need to adapt alternate schemes. II) The equations of motion are time-reversible, thus allowing not only the calculation of future configurations but also past configurations. III) It is a symplectic integration, meaning the probability density of phase space is conserved as time progresses as dictated by Liouville's theorem and is one of the fundamental properties in the description of statistical mechanics as discussed in the introduction. [42]

To propagate the time in molecular dynamics, we thus need to know the acceleration of the individual particles, which according to the velocity Verlet integration scheme (Eq. 4.2) depends on the positions of the particles, with Newton's second law of motion relating the net force acting on a particle to the particle's acceleration. This force is also related to the variation in potential energy $\mathcal{V}$ with respect to the distance to the object exerting the force,

$$\vec{F} = -\nabla\mathcal{V}(\vec{q}) = -\left(\frac{\partial\mathcal{V}(\vec{q})}{\partial x}, \frac{\partial\mathcal{V}(\vec{q})}{\partial y}, \frac{\partial\mathcal{V}(\vec{q})}{\partial z}\right). \tag{4.3}$$

Here we have adapted a three-dimensional notation, emphasizing the need to differentiate with respect to all spacial dimensions. The force evaluation creates a limitation as to which systems can be studied using molecular dynamics as every energy function employed must be differentiable. Consequently, potentials such as hard-spheres and square-well potentials are not compatible with molecular dynamics simulations.

Using Newton's equations of motion, we can write up how the energy will vary with time, given the fact that the total energy (the Hamiltonian) can be written as a double sum with the first involving the individual particles in the system and the second sum being the kinetic energy, $\mathcal{K}$, and potential energy, $\mathcal{V}$, we can write

$$\frac{\partial\mathcal{H}}{\partial t} = \frac{\partial}{\partial t}\sum_{i=1}^{N}(\mathcal{K}_i + \mathcal{V}_i) = \sum_{i=1}^{N}\frac{\partial}{\partial t}(\mathcal{K}_i + \mathcal{V}_i) = \sum_{i=1}^{N}\frac{\partial}{\partial t}\left(\frac{m_i v_i^2}{2} + \mathcal{V}_i\right). \tag{4.4}$$

Rewriting the first term on the right hand side as $\frac{dv^2}{dt} = 2v\frac{\partial v}{\partial t}$, and the second term as:

$\frac{\partial \mathcal{V}}{\partial t} = \frac{\partial \mathcal{V}}{\partial q}\frac{\partial q}{\partial t}$, both using the chain rule of derivatives, the above expression then becomes

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial t} &= \sum_{i=1}^{N} m_i v_i \frac{\partial v_i}{\partial t} + \sum_{i=1}^{N} \frac{\partial \mathcal{V}_i}{\partial q_i}\frac{\partial q_i}{\partial t} = \sum_{i=1}^{N} m_i a_i v_i + \sum_{i=1}^{N} \frac{\partial \mathcal{V}_i}{\partial q_i} v_i \\ &= \sum_{i=1}^{N} F_i v_i - \sum_{i=1}^{N} F_i v_i = 0.\end{aligned} \tag{4.5}$$

From Eq. 4.5 we can conclude Newton's equations of motion have zero variation in the total energy with respect to time and thus the system possesses constant energy as the sampling occurs, revealing molecular dynamics to nativity sampling the microcanonical (*NVE*) ensemble. While the total energy in practice does vary, this effect arises from the non-infinitesimal time increment ($\Delta t$) when using Eq. 4.2, and thus depends on the choice of $\Delta t$ and system being simulated.

Due to MD natively sampling the microcanonical ensemble by solving Newton's equations of motion, to sample the canonical ensemble the need for a method of controlling the energy dissipation between the system and surroundings is required to obtain constant temperature rather than constant energy. These methods for controlling the temperature are called *thermostats* and are commonly applied in molecular dynamics simulations today, to mimic experimental conditions. To calculate the temperature in a molecular system, one commonly utilizes the *law of equipartitioning* stating that at equilibrium the energy of a system is equally partitioned to the various microscopic modes of motion and configurations in terms of the thermal energy $k_{\mathrm{B}}T$ times a constant.[53] This has the consequence that one may calculate the instantaneous temperature for a system at any given time using configurational or dynamical quantities of the phase space.[53,104,109] For practical purposes, the most commonly utilized microscopic mode of motion utilized for the calculation of temperatures, is the velocities of the particles in the system, which in turn is related to the average kinetic energy. In particular, one consequence of the equipartition theorem is that any energy mode that depends quadratic on a phase-space variable possesses $\frac{1}{2}k_B T$ energy per degree of freedom. Due to molecular dynamics retaining information on the velocities, we can write the expression

$$T(t) = \frac{2\mathcal{K}}{3Nk_B} = \frac{1}{3Nk_B}\sum_{i=1}^{N} m_i |\vec{v_i}(t)|^2. \tag{4.6}$$

In Eq. 4.6 $T(t)$ is known as the instantaneous kinetic temperature and allows the calculation of the temperature at any point during the molecular dynamics simulation. To properly sample the canonical ensemble the time-averaged temperature must equal the desired temperature: $\langle T(t) \rangle = T$. To achieve this, one can modify the equations of motion to ensure the time-averaged ensemble of instantaneous kinetic temperatures has the correct

mean. The simplest and most naive approach would be to scale the velocity of all the individual particles in Eq. 4.2 by the factor $\sqrt{T/T(t)}$ thus achieving constant temperature after every iteration. However, despite the method providing a trajectory at a constant temperature, it turns out this approach does not generate the canonical ensemble, due to complete removal of fluctuations in the instantaneous temperature, but instead samples the so-called isokinetic ensemble. [1] Using the central limit theorem for the total kinetic energy of a system containing $N$ particles we can derive the variance of the probability distribution is given by $2T^2/(3N)$, thus revealing any system, not in the thermodynamic limit ($N \to \infty$) will display fluctuations in temperature. To circumvent the strong coupling in exchange of energy between system and surroundings as it was found for the isokinetic ensemble Berendsen and coworkers[9] scaled the velocities in a time-dependent manner, such that the rate of change in temperature is proportional to the difference in temperature

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = \frac{1}{\tau}[T - T(t)]. \tag{4.7}$$

In Eq. 4.7 $\tau$ is a coupling parameter mimicking how strong the system is coupled to the heat bath, with the extrema for the coupling parameter being $\tau = \Delta t$ yielding the equations of motion for the isokinetic ensemble and $\tau = \infty$ yielding the equations of motion for the microcanonical ensemble (Eq. 4.2). Morishita revealed the phase-space distribution using the Berendsen thermostat can be generalized to the so-called Berendsen ensemble, which in addition to the natural variables of $NVT$ also depends on the choice of $\tau$, and is an interpolation between the microcanonical ensemble and isokinetic ensemble, with the canonical ensemble being a subspace of the Berendsen ensemble within the interpolation. [84]

Many other thermostats exist, each having its pros and cons, with the main utilized methods mentioned in Tbl. 4.1. The methods are grouped according to their methodology in achieving constant temperature. For example, the previously mentioned methods all utilize velocity rescaling in a more or less deterministic form, however, some methods employ stochastic forces and particle collisions, while finally, the last group of methods employs an extended Lagrangian methodology in which the system is coupled to artificial particles via their coordinates and velocities.

To obtain constant pressure one relies on the usage of a *barostat*, which like thermostats alters the equations of motion. The pressure at any time $t$ from molecular dynamics simulations is commonly estimated from the viral pressure equation

$$P(t) = \frac{NkT}{V} - \frac{1}{6V}\sum_{i=0}^{N}\sum_{i\neq j}^{N}\vec{q}_{ij}\cdot\vec{F}_{ij}. \tag{4.8}$$

[1] This approach only works for sufficiently small time step to ensure the discontinuities in momentum is not too great. [13] To accommodate this problem the Gaussian thermostat is applied which modifies Newton's equations of motions to preserve the kinetic energy via Gauss's principle of least constraint.

**Table 4.1:** Thermostats within molecular dynamics grouped according to the thermostat's method of achieving constant temperature. The references listed are recommended literature on the theory behind the individual thermostats.

| Method | Thermostat | References |
|---|---|---|
| Velocity Rescaling | Isokinetic thermostat | 35, 36, 147 |
| | Berendsen thermostat | 9, 84 |
| | Canonical velocity rescaling (Bussi thermostat) | 17–19 |
| Stochastic Forces | Andersen thermostat | 2 |
| | Langevin thermostat | 37, 62 |
| | Dissipative particle dynamics | 88, 124 |
| Extended Lagrangian | Nosé-Hoover thermostat | 48, 91 |
| | Nosé-Hoover chains | 73 |

In Eq. 4.8 the first term is the contribution arising from ideal particles interacting with the system boundaries, while the second term accounts for the interactions between particles with the scalar product between the position vector and force vector known as the viral. As with thermostats, to achieve successful barostating the ensemble average of the instantaneous pressure must equal the desired pressure $\langle P(t) \rangle = P$ with correct fluctuations. Consequently, as with thermostats, simple re-scaling methods of the pressure to obtain constant pressure after every time step or in a time-dependent manner (equivalent to Eq. 4.7) as applied in the Berendsen barostat,[9] is not recommended due to the sampled ensemble not being well defined due to suppression of fluctuations. Another option is to change the equations of motion by the addition of an additional degree of freedom as utilized in the Andersen barostat[2], thus behaving as if an isotropic piston is acting on the system. However, while this barostat does sample the correct ensemble, it is restricted to isotropic pressure regulation. The Parrinello-Rahman barostat[94] addressed the support for anisotropic scaling and like the Andersen barostat samples the correct ensemble, with the extra property of allowing to change simulation box shape, which can be highly useful in the simulation of solids. One downside of the Parrinello-Rahman barostat is for the equations of motion to only hold in the thermodynamic limit. Consequently, the Martyna-Tuckerman-Tobias-Klein barostat[74,75] altered the equations of motion to allow for the correct sampling of finite-sized systems and is to date perhaps the best generally applicable barostat that preserves dynamic fluctuations in molecular dynamics.[14] The last option is the usage of a Monte Carlo barostat, which can be proven to also sample the correct ensemble.[23] While Monte Carlo barostating is highly efficient and simple to implement, it comes with the cost that the dynamics are destroyed as the kinetic regime of the phase space is omitted in Monte Carlo simulations.

## 4.2 Langevin Dynamics

From molecular dynamics, we have established that it is hard to maintain the true dynamics of a system, while simultaneously being able to correctly sample the canonical ensemble. From Tbl. 4.1 the velocity re-scaling and extended Lagrangian methods were presented as fully deterministic methods while attempting to sample an ensemble of constant temperature. In the meantime, the usage of stochastic forces, despite the possibility of disturbing the dynamics, has proven highly successful to sample the canonical ensemble. The perhaps most impacting example is that of Langevin dynamics and the science around it which has been a contributing factor in shaping our current view on the world today.

In 1827 the botanist Robert Brown found that pollen grain particles immersed in water under a microscope were in constant motion following irregular paths.[15] Despite this observation was already done in 1785 by Jan Ingenhousz from observing the behavior of coal dust on a surface of alcohol,[96] this type of motion came to be known as *Brownian motion*. Until the 1900s this motion was mostly worked out using *random walks*, however, in 1905 Albert Einstein[33] and 1906 Marian Smoluchowski[139] showed this irregular motion originated from the collision between larger particles (pollen grain) and much smaller particles (water) in a heat bath, thus indefinitely proving the existence of atoms. While these two findings provided a conceptional leap to understand the origin of the motion, the theory was limited to a qualitative description of the motion, until the development of stochastic differential equations that can accurately capture the movement of Brownian particles by Paul Langevin in 1908,[64] which came to be known as Langevin dynamics. The Langevin equations of motion are given by[28,62]

$$\begin{aligned} \frac{\mathrm{d}\vec{q}}{\mathrm{d}t} &= \vec{v} \\ \frac{\mathrm{d}\vec{v}}{\mathrm{d}t} &= \frac{-\nabla\mathcal{V}(\vec{q})}{m} - \frac{\gamma}{m}\vec{v} + \frac{\sqrt{2\gamma k_B T}}{m}\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}, \end{aligned} \tag{4.9}$$

where $\gamma$ is named the *friction coefficient* having the unit of mass times inverse time, and $\vec{W}$ is the Wiener process responsible for the stochastic character of the time evolution with the process drawing random numbers from a specific probability distribution, in this case, the Gaussian distribution. In other literature, the acceleration component of the Langevin equation may also be expressed using white noise, $\eta(t)$, which is simply the time derivative of the Wiener process having the unit of square root time over time. The Langevin equations of motion can be found to be highly equivalent to Newton's equations of motion, however with the difference being the addition of the two last terms in the expression for acceleration. These two terms are the consequence of the fluctuation-dissipation theorem, in which kinetic energy possessed by solute molecules is converted into thermal energy via drag forces due to the solvent constituting the second term of Eq. 4.9 being the Stokes' drag. The third term is a stochastic term arising from the fact the solvent molecules are

in constant motion thus colliding with the solute molecules, yielding kinetic energy to the solute molecules. The equations of motion for Langevin dynamics can thus be given the physical interpretation that particles are immersed in a bath of many small and light particles interacting via collisions or weaker interactions than those described by the potential $U$. This effectively means we can approximate the forces arising from the bulk without the need to direct simulate them, and exactly this idea should sound appealing to any modeler. Finally, it is worth noticing the Langevin equations of motion (Eq. 4.9) can like the Newtonian equations of motion (Eq. 4.2) be found to obey the Markov property, meaning the future evolution of the system only depends on its current state and not prior history. The application of this will be discussed further in chapter 4.4.

One of the key advantages of the usage of Langevin dynamics is the possibility to prove the stationary distribution generated by solving the equations of motion is the canonical distribution function. This can be derived using the time-dependent probability distribution function for Langevin dynamics is given by the *Fokker-Planck equation*. In specific we can write the configurational part of the Langevin equation as

$$\frac{\mathrm{d}\vec{q}}{\mathrm{d}t} = -\frac{1}{\gamma}\nabla\mathcal{V}(\vec{q}) + \sqrt{\frac{2k_BT}{\gamma}}\eta(t). \tag{4.10}$$

The corresponding Fokker-Planck equation can be written as

$$\frac{\partial\rho(\vec{q},t)}{\partial t} = \frac{1}{\gamma}\frac{\partial}{\partial\vec{q}}\left[\nabla\mathcal{V}(\vec{q})\rho(\vec{q},t)\right] + \frac{k_BT}{\gamma}\frac{\partial^2\rho(\vec{q},t)}{\partial\vec{q}^2}. \tag{4.11}$$

To find the stationary distribution, we set Eq. 4.11 to zero, furthermore, we can factor out one of the derivatives with respect to the position to find the expression for the *probability current* which must also yield zero when the distribution is stationary, thus leaving the differential equation

$$\frac{1}{\gamma}\nabla\mathcal{V}(\vec{q})\rho(\vec{q}) + \frac{k_BT}{\gamma}\frac{\partial\rho(\vec{q})}{\partial\vec{q}} = 0, \tag{4.12}$$

which can be solved to

$$\rho(\vec{q}) \propto e^{-\beta\mathcal{V}(\vec{q})}, \tag{4.13}$$

which is the Boltzmann distribution. Similarly, we can write the kinetic part of the Langevin equation as

$$\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = \frac{1}{m}\left(-\gamma\vec{v} + \frac{\sqrt{2\gamma k_BT}}{m}\eta(t)\right), \tag{4.14}$$

with the corresponding Fokker-Planck equation

$$\frac{\partial\rho(\vec{v},t)}{\partial t} = \frac{1}{m}\frac{\partial}{\partial\vec{v}}\left(\gamma\rho(\vec{v},t)\vec{v} + \frac{k_BT\gamma}{m}\frac{\partial\rho(\vec{v},t)}{\partial\vec{v}}\right). \tag{4.15}$$

The stationary distribution is ones more found by setting the time variation of the probability distribution function and the probability current equal to zero, thus finding

$$\frac{\partial \rho(\vec{v})}{\partial \vec{v}} = -\frac{m}{k_B T}\rho(\vec{v})\vec{v} \Rightarrow \rho(\vec{v}) \propto e^{\frac{-m\vec{v}^2}{2k_B T}}, \tag{4.16}$$

which is the well-known Maxwell-Boltzmann distribution.

Having shown that Langevin dynamics yields the Boltzmann distribution, as its stationary distribution, we now face the issue of integrating the Langevin equations of motion to construct a sampling iteration scheme as it was Newtonian dynamics in Eq. 4.2. Leimkuhler, Matthews, and contemporaries proposed the Langevin equation could with advantage be divided into separate parts, in which we decompose the Hamiltonian dynamics into its velocity and positional component, and the Ornstein-Uhlenbeck component as a separate term.[62,79] The Langevin equations on differential form can thus be written as

$$\mathrm{d}\begin{bmatrix}\vec{q}\\ \vec{v}\end{bmatrix} = \underbrace{\begin{bmatrix}\vec{v}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathrm{A}} + \underbrace{\begin{bmatrix}0\\ -\nabla\mathcal{V}(\vec{q})m^{-1}\end{bmatrix}\mathrm{d}t}_{\mathrm{B}} + \underbrace{\begin{bmatrix}0\\ -\gamma'\vec{v} + \sqrt{2k_B T\gamma'}\frac{1}{\sqrt{m}}\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}\end{bmatrix}\mathrm{d}t}_{\mathrm{O}}. \tag{4.17}$$

Note in Eq. 4.17 that the friction coefficient $\gamma$ has now been replaced by the *collision rate*, $\gamma'$ ($\gamma = m\gamma'$), being a frequency and therefore having the units of inverse time, which is a more common input in molecular dynamics software packages. In Eq. 4.17, the A step can be recognized as a "drift" step, the B step can be recognized as a "kick" step, and O is a new step we will term the "fluctuate" step due to its stochastic nature. The individual differential equations for the three steps may be solved exactly to yield the changes in position and velocity upon an increment in time step. We may write a single particle's position in the position-velocity phase space at time $t + \Delta t$ as

$$\begin{aligned}
(\vec{q}(t+\Delta t), \vec{v}(t+\Delta t))^{\mathrm{A}} &= (\vec{q} + \Delta t\vec{v}, \vec{v})\\
(\vec{q}(t+\Delta t), \vec{v}(t+\Delta t))^{\mathrm{B}} &= \left(\vec{q}, \vec{v} - \Delta t\nabla\mathcal{V}(\vec{q})\frac{1}{m}\right)\\
(\vec{q}(t+\Delta t), \vec{v}(t+\Delta t))^{\mathrm{O}} &= \left(\vec{q}, e^{-\gamma'\Delta t}\vec{v} + \sqrt{k_B T(1-e^{-2\gamma'\Delta t})}\frac{1}{\sqrt{m}}\mathcal{N}(0,1)\right),
\end{aligned} \tag{4.18}$$

where $\mathcal{N}(0,1)$ is a random number from the normal distribution (Gaussian distribution with zero mean and variance one). Using these individual steps we can construct families of numerical integrators based on the sequence of the letters "ABO". For the method to be consistent in updating the positions and velocities to the correct time-step upon reaching the end of the string, we require that if a letter appears $k$ times in the method's string, the individual updates for the specific operations should use the time step $\frac{\Delta t}{k}$. Among the schemes utilized in this work are the integration schemes $[\![\mathrm{BAOAB}]\!]$, which can also

be named "symmetric Langevin velocity-Verlet" (used with openMM[105]), and ⟦BBAOA⟧, which can be named "Langevin leapfrog" (used with Amber[32]). In the limit of an infinitely small time step, the two methods should yield identical trajectories. However, for practical purposes, the utilization of a finite time step will come with an associated error, with the error being highly dependent on the choice of integration scheme both in terms of context and magnitude. In regards to context, the most commonly encountered problem is the issue of maintaining a constant temperature. As previously explained arbitrary phase space vector fields may be utilized to relate microscopic details to the instantaneous temperature of the system. In particular, it has been shown the ⟦BAOAB⟧ scheme yields highly accurate temperature distributions when utilizing configurational temperatures, while the kinetic temperature distribution suffers. On the other hand, the ⟦BBAOA⟧ scheme yields highly accurate temperature distributions when utilizing kinetic temperatures, due to the leapfrog nature of the scheme, while the configurational temperature distribution is off.[38,62,63] Given this binary relationship of methods excelling in accuracy in either the configurational or kinetic space for the calculation of molecular properties, the most logical choice of Langevin integration scheme should reflect the desired property to be calculated. In particular for properties that depend only on the configuration space and free energies are highly recommended the ⟦BAOAB⟧, while kinetic properties like time-correlation functions, can be better computed using an integrator which balances the errors from both the configurational and kinetic space.

## 4.3 Markov Chain Monte Carlo Simulations

Up to this point, we have discussed methods that preserve the dynamical information of the system, thus yielding a trajectory to which at any time, the position and velocity of each particle contained within the system can be determined, and can be used to calculate ensemble averages of molecular properties. However, one may address the question: "why use trajectories to calculate ensemble averages?". Since our approach to obtaining ensemble averages is the computation of expectation values for variables for a given probability density, it has nothing to do with dynamics. While for example solving the Langevin equations of motion will converge to a stationary distribution, namely the Boltzmann distribution, one should be able to choose any scheme which will converge to the Boltzmann distribution, regardless of the preservation of dynamical information. Since the Hamiltonian can be decomposed into kinetic and potential energy independent terms and the canonical partition function to the kinetic energy is equal to the partition function of an ideal gas at constant temperature and thus can be solved analytically, one can argue the need for dynamics is irrelevant. The dimensional removal of the $3N$-momenta space is with advantage utilized Markov Chain Monte Carlo (MCMC) simulations.

Markov Chain Monte Carlo simulations utilize a random weighted walk in the configu-

rational subspace of phase space. By this description, MCMC can be decomposed such that the weighting of space is conducted within the theory of Markov chains and the random walk/sampling will be conducted by Monte Carlo experiments. A Markov chain is as described a model in which a finite or infinite number of states obeys the Markov property, i.e. the transition between states only depends on the currently acquired state and is thus independent of any past or future outcome. Instead, the future and past events are determined by a transition matrix containing within it the probabilities of going from one microstate to another or itself. Using the probabilities is donated within the range [0,1], the transition matrix must fulfill the condition

$$\sum_{j=1} T_{i,j} = 1, \forall j \in \Gamma_q, \tag{4.19}$$

where $T_{i,j}$ is the transition matrix going from microstate $i$ to $j$, and all possible states $i, j$ belongs to the configurational phase space $\Gamma_q$. For the majority of Markov Chains, we necessitate the requirement of *detailed balance*,[2] implying the Markov chain to be fully reversible, meaning the net flow between any two states, whether configurational or microscopic, must yield zero. One may write the detailed balance criterion as

$$\pi_i T_{i,j} = \pi_j T_{j,i}. \tag{4.20}$$

Where $\pi$ is the probability density for a given state. The detailed balance requirement is in many ways equivalent to the concept of systems being in equilibrium, as probability density functions can be exchanged with populations of chemical species, and the transition matrix describing the transformation of one chemical species into another can be exchanged for rate constants and equilibrium constants.

For MCMC simulations the most commonly utilized algorithm to generate configurations with Boltzmann distributed probabilities is the Metropolis algorithm,[81] in which we choose the transition matrix should process two stochastic kernels, the first being the random choice of selecting a specific particle and Monte Carlo move $\alpha(j|i)$, and the second being the acceptance of the move $\mathrm{acc}(j|i)$. The probability of multiple independent events occurring is given by the product of the individual independent events occurring, hence the transition matrix is given by

$$T_{i,j} = \alpha(j|i)\ \mathrm{acc}(j|i). \tag{4.21}$$

Utilizing the above expression in the expression for detailed balance, and utilizing the ratio between two states of different potential energy is given by the Boltzmann distribution we find any newly generated configuration should be accepted with probability

$$\mathrm{acc}(j|i) = \min\left(1, e^{-\beta(\mathcal{V}_j - \mathcal{V}_i)}\right). \tag{4.22}$$

[2] The requirement of detailed balance is too strict,[71] and one may do Monte Carlo simulations under weaker balance conditions, however, the maintenance of balance compared to detailed balance is more difficult to access.[40]

However, Eq. 4.22 merely favors configurations of lower energies. To check if the configuration $j$ can exist in thermal equilibrium with $i$ we generate a random number from the continuous uniform distribution, $\mathcal{U}(0,1)$, and accept if $\mathcal{U}(0,1) < \mathrm{acc}(j|i)$. The Metropolis Monte Carlo algorithm can be summarized as:

**The Metropolis Monte Carlo Algorithm**

- Generate initial configuration with a potential energy $\mathcal{V}_i$.
- Until convergence has been obtained:
  - Generate a trial configuration with a potential energy $\mathcal{V}_j$.
  - **If** $\mathcal{V}_j < \mathcal{V}_i$: Accept.
  - **Else if** $\mathcal{U}(0,1) < \exp(-\beta[\mathcal{V}_j - \mathcal{V}_i])$: Accept.
  - **Else**: Reject.
  - Sample configuration for molecular properties.
- Terminate simulation.

Accept: Adapt trial configuration as new initial configuration for future energy-difference evaluations.
Reject: Continue with the initial configuration for future energy-difference evaluations.

We have now illustrated two schemes, at which the canonical ensembles can be sampled using either molecular dynamics or Monte Carlo simulations. While we will in the upcoming discuss the possibility to draw on the strength of both (chapter. 4.4), an amusing discussion is usually the performance of molecular dynamics versus Monte Carlo simulations to most effectively sample the ensemble distribution. One of the advantages of Monte Carlo is it is inherently faster per iteration than molecular dynamics due to the unnecessity of having to calculate a force matrix, but only the potential energy, being a scalar value, and by the same argument is more adaptive in terms of potentials, as continuous and differentiable potential energy functions are not a requirement. The main disadvantage of Monte Carlo is particularly related to one point: the generation of trial configurations. For highly correlated systems, as it is for example seen for systems being modeled using steep potentials, the Monte Carlo method can be subject to a high rejection rate. This is particularly a problem when many modes of motion are correlated and one attempts to only update one mode by a Monte Carlo move, thus yielding high energies. Additionally Monte Carlo simulations are unguided in their search for high probability density in phase space, forced to randomly search the phase space which can be highly complicated with steep barriers in phase space, thus causing Monte Carlo simulations to spend a lot of time yielding no new information. On the other hand, molecular dynamics utilize the forces to guide their movement in phase

space and are thus often superior to Monte Carlo simulations in their most "crude form" with non-optimized parameters.

## 4.4 Combined Molecular Simulation Schemes

MC simulations and MD simulations are fundamentally different in their approach to generate the desired thermodynamic ensemble. While molecular dynamics generates a single long trajectory of the system through phase space, MC typically samples the configurational space via a nonphysical path. However, inherently in the difference of the methods to generate the desired thermodynamic ensemble the difficulties associated with the method are also different. In particular, the sampling of multiple free energy minima separated by energy barriers hinders the sampling by molecular dynamics, with the rate constant between the minima being proportional to the exponential magnitude of the energetic barrier as given from transition state theory and the Arrhenius equation. While MC however can overcome such problems, it relies on the simulator having an initial knowledge about the system to be simulated such that MC moves can be designed to effectively sample the configurational space, as it could in otherwise worst consequence yield universal rejection of the MC moves and thus provide no useful results what so ever. As of consequence, it would be beneficial to combine the strengths of the two simulation methods; molecular dynamics' inherent sampling of the phase space by simply following the forces governing the molecules and MC's possibility to be highly adaptive in the jumping between phase space points. Methods utilizing both MD and MC are commonly referred to as *combined MDMC methods* and are much more commonly utilized, than perhaps first anticipated. The perhaps most common usages include the sampling of exotic ensembles such as the grand canonical ensemble ($\mu VT$), the isothermal-isobaric ensemble ($NPV$), and the semi-grand isothermal-isobaric ensemble ($\Delta\mu PT$), but also enhanced sampling methods such as replica exchange and adapted MD/MC protocols in which only some particles are moved by MD and others by MC. However, to understand and appreciate these methods we will first discuss the fundamental property required to combine MD and MC.

In chapter 4.3 we discussed how MC simulations utilize a Markov process which is a stochastic process that satisfied the condition, the system in its current state is independent of future and past states for the generation of configurational states. While MD simulations mimic the natural dynamics of molecules, MD is, in fact, *also* a Markov process however involving both deterministic components, like the gradient of the chosen potential energy function and drag forces, and stochastic components as it could, for example, be achieved from thermostating or the thermal noise from a Wiener process[3] as found in Langevin dynamics. Even in the case of Newtonian dynamics with completely deterministic description

[3] Which is also memoryless as discussed in section 4.2.

in the canonical ensemble as it can be achieved via Nose-Hoover (chains) thermostating or Berendsen thermostating the stochastic probability matrix is simply a $\delta$-function,[31] given the requirement the equations of motion are integrated to be reversible. Consequently, one might be tempted to say that everything is Markovian, or at least "effectively Markovian" as long as the full phase space is treated characteristic for the given method, with molecular dynamics requiring both treatments of position and momenta to appear Markovian.[153]

Three classes of approaches in their combination of MD and MC simulations have been identified[87] being: *mixed MC/MD* characterized by some atoms are moved by MC and others by MD, *hybrid MC/MD* characterized by the unification of the MC and MD algorithm to generate the system, and finally, *sequential MC/MD* in which one apply the algorithms independently but in an alternate, sequential fashion. While the mixed MC/MD scheme is not so commonly utilized in the molecular simulation of liquids, among others due to the increased computational cost of mixing[60] and little gain in terms of sampling for dense systems,[89] the sequential MC/MD scheme is now commonly utilized methods and hybrid MC/MD scheme is fast gaining ground in terms of development and availability. For the hybrid MC/MD scheme one of the latest added methods is the so-called non-equilibrium candidate Monte Carlo move,[89] which can be considered a "meta MC"-move in the sense this move takes a perturbation kernel and propagation kernel as arguments. The perturbation kernel may be any perturbation that brings the system out of equilibrium and can include changes in configurational space, number of particles, composition, system dimensions, etc. while the propagation kernel is commonly taken to be molecular dynamics, Langevin dynamics, or Metropolis MC. The main difference from a normal MC move is thus though a finite process to evaluate if the perturbation generated can be accepted by using the non-equilibrium work done by the propagation kernel in the acceptance criteria, rather than the instantaneous energy difference as used in ordinary Metropolis MC simulations. While this move is more expensive per iteration, it has been shown to yield a high acceptance rate for moves that may otherwise face a high rejection rate by ordinary MC. An example of such a move includes the configurational sampling of a bi-stable dimer in a WCA solvent, in which the WCA solvent is effectively preventing the transition between the two minima, due to the need for first displacing the solvent, which can not be overcome by ordinary MC and molecular dynamics.[89]

Among the now commonly applied methods today using sequential MC/MD includes Monte Carlo barostating of molecular dynamics simulations, in which a molecular dynamics simulation is periodically interrupted by attempting a volume move, thus increasing or decrease the simulation box dimensions and molecular coordinates, thus sampling the isothermal-isobaric ensemble ($NPT$) when the molecular dynamics simulation is accompanied with a thermostat[4,23]. Unfortunately, Monte Carlo barostating does not allow the sampling of the isoenthalpic-isobaric ensemble ($NPH$), due to the inherent canonical distribution of the Metropolis Monte Carlo algorithm, thus effectively preventing the

scaling of volume at constant energy.

Another common usage of the sequential MC/MD scheme is the discrete constant pH molecular dynamics (CpHMD) method[16,82] and highly related discrete constant redox potential molecular dynamics method. The two methods are highly related, due to the mathematical similarity between the Henderson-Hasselbalch equation, which is applied to acid-base reactions, and the Nernst equation, which is applied in electrochemistry. The main difference between the two is usage of the decimal logarithm in traditional acid-base chemistry to express the proton activity/concentration and the chemical potential of protons remaining fixed in oppose to constant redox potential molecular dynamics having the chemical potential of electrons fixed[27]. Thus we will henceforth only discuss CpHMD and remind the reader interested in redox reactions and molecular dynamics everything is equally applicable to redox reactions, but slightly modified.

### 4.4.1 Constant pH Molecular Dynamics

Up to the development of constant pH molecular dynamics (CpHMD), conventional molecular dynamics was limited to the utilization of fixed protonation states for titratable acids and bases. This possesses a series of drawbacks, the first being the actual assignment of protonation states, which requires intrinsic knowledge on the p$K_a$ value of the individual acids and bases. While this problem is majorly simplified for acids and bases in an aqueous environment, the problem is on the other hand tremendously complicated when the acids and bases are found in a highly heterogeneous environment, such as in titratable residues in proteins[93] or phosphatidic acids lipid bilayers.[116] The second drawback is; if the pH is near or equal to the intrinsic p$K_a$ value for the acids or bases, there is no single protonation state to represent the ensemble of protonation states appropriate at the pH of interest, with can be further complicated when different conformational states may be characterized by different p$K_a$ values, thus uncoupling the dynamics from the acid-base equilibrium. To accommodate these drawbacks the CpHMD method is taking advantage of the efficient conformational sampling by molecular dynamics, while the sampling of Boltzmann distributed protonation states is done by MC.[16,82]

The breaking of a chemical bond is fundamentally a quantum mechanical phenomenon thus unable to be accurately described by classical methods. As of consequence, in order to describe the free energy of protonation we utilize a transfer model of model compounds from an aqueous solution to the system of interest (see Fig. 3.4). Taking the deprotonated state as the ground state, the protonation free energy is given by

$$\Delta G_{\text{protonation}} = k_B T \ln 10(\text{pH} - \text{p}K_{a,\text{ref}}) + \Delta G_{\text{ele}} - \Delta G_{\text{ele,ref}}, \tag{4.23}$$

where pH is the acidity of the system, $\text{p}K_{a,\text{ref}}$ is the p$K_a$ value of the reference compound in aqueous solution determined by experimental methods, $\Delta G_{\text{ele}}$ is the difference

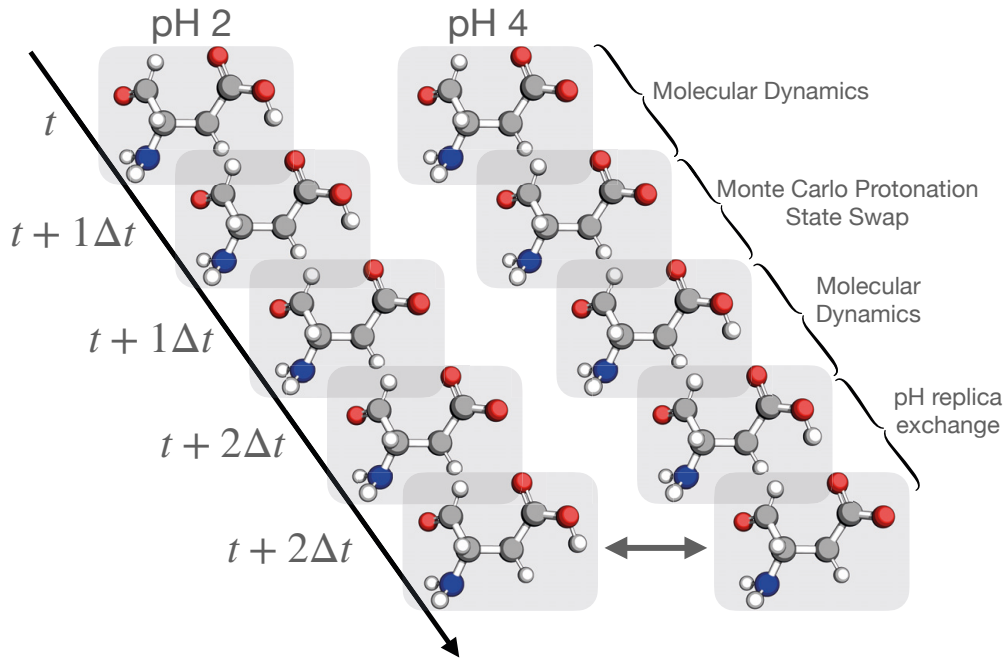



**Figure 4.2:** Illustration of the discrete constant pH molecular dynamics method with pH replica exchange. Independent simulations with different pH are conducted which samples the conformational space by molecular dynamics using a finite time step $\Delta t$, at periodic time intervals the molecular dynamics sampling is stopped and a Monte Carlo move is performed which can either be a protonation state swap move or a pH replica exchange move. The protonation state swap move is mainly to assist in the sampling of protonation space, while the pH replica exchange move mainly enhances the pH-dependent conformational sampling.

in Hamiltonian between the proposed protonation state and current protonation state, while $\Delta G_{\mathrm{ele,ref}}$ is a precomputed quantity, ensuring the protonation free energy equals zero when $\mathrm{pH} = \mathrm{p}K_{a,\mathrm{ref}}$ for the model compound in water and thus depends on the choice of force field and solvation model.[82] The model compound is usually taken to be the N-methylated and C-acetylated amino acid residue derivatives, and the value is numerically estimated by free energy methods, with the potential need of minor manual tweaks to ensure an equal sampling of the deprotonated and protonated state at $\mathrm{pH} = \mathrm{p}K_{a,\mathrm{ref}}$. In the Amber implementation of CpHMD, implicit solvation is being utilized for the protonation state evaluation, in particular generalized Born solvation[6] with the possibility of including implicit Debye–Hückel based salt.[125]

The practical aspect of the CpHMD protocol is thus to conduct normal MD simulations with fixed protonation states accompanied by periodic MC moves attempting to change the protonation state of random titratable residues. In the Amber implementation, the structural sampling can be conducted using either implicit or explicit solvent, while the protonation state sampling is limited to implicit solvent only. The limitation of not being able to use explicit solvent for the protonation state sampling are mainly due to two issues;

first is the excluded volume exerted by the proton upon insertion, causing potential particle overlap with nearby solvent molecules and second is the solvent reorganization energy due to the creation or removal of electric charge. The naive solution[4] is simply just to temporarily substitute the explicit solvent with implicit solvent, thus removing all the explicit solvent from the simulation upon conducting a MC protonation state move and reinserting the explicit solvent upon continuing the sampling by molecular dynamics. To prevent huge energies upon a successful protonation state move, the solvent reinsertion is followed by a solvent relaxation by some finite amount of time, while upon an unsuccessful protonation state move the solvent is simply reinserted. Finally, to maintain charge neutrality within the system, due to free protons not being treated explicitly, a random water molecule is being transformed into an explicit monovalent ion thus neutralizing the charge causing by the proton insertion/removal.

To capture strongly coupled titration equilibria between protons close in space, multi-site protonation state moves i.e. the simultaneous change in multiple protonation states can with great advantage be adapted over subsequent single protonation state moves.[10] In the Amber implementation of CpHMD, residues can engage in multi-site protonation state moves if the titrating protons are within 2 Å of each other with the multi-site move having a 0.25 weight compared to a single-site move. The multi-site move allows for proton transfer between titratable residues involved in hydrogen bonding, with the protein transfer experiencing rejection only from single-site moves, due to the need of first breaking the hydrogen bond causing a high energy penalty.

In addition to the MC protonation state move, which is one of the main elements of discrete CpHMD, another MC move can with great advantage be utilized in parallel with the protonation state move, namely pH replica exchange. The conformational sampling was by Mongan *et al.* identified for lysozyme to be the limiting step to achieve convergence of p$K_a$ values,[82] however utilizing pH replica exchange the sampling was improved.[51,128] Due to the ratio of transition probabilities not depending on the Hamiltonian of the system, but the difference in pH between two systems and the number of titratable protons within the system, the replica-exchange Monte Carlo is essentially a cheap method to gain enhanced sampling, as energy and force evaluations are commonly the most time-consuming process for larger systems. Within the sequential MD/MC scheme the two MC moves should not be attempted simultaneously, but instead choose a Monte Carlo move at random during the MC iteration, to keep the replica-exchange Markov chain and protonation state Markov chain uncoupled.[51]

[4]A discussion of the problems and the potential solutions associated with this method can be found in 5.2.

# Chapter 5

# Summary and Reflections on Thesis Work

*Knowledge is knowing a tomato is a fruit; wisdom is not putting it in a fruit salad.*
— Miles Kington

In this chapter, we will overview the main findings and conclusions from the different papers, in addition to reflections and considerations done post-publication. Additionally, we will also discuss methods and associated theories for optimizing the solubility of molecular matter in a desired phase or state.

## 5.1 Intrinsic & Extrinsic Factors for Improving Solubility

In order to make a successful drug, one mainly faces two challenges:

1. The drug possesses the chemical properties necessary to achieve the desired biological effect, such as acting as inhibitors for proteins possessing correct molecular geometry and coordination chemistry, or therapeutic proteins possessing the correct fold to engage in the regulatory process targeted.
2. The drug needs to be successfully delivered to the key locations required for regulation while remaining solubility not to become excluded from the body or even becoming toxic.

While the majority of the research in this work has focused on issue number two, the change in proteins' structural stability by changes in the solvent composition has been ex-

plored using a very simple, yet highly enlightening, transfer model presented in chapter 2.2.1. To tackle issue number two, we have been exploring the governing principles for controlling the aggregation of molecular matter, characterized by the preference for solute molecules' interactions with themself, rather than the surrounding solvent, causing the solute to become biologically inactive or toxic. Additionally, we have also explored the governing principles for controlling the partitioning of molecular matter in various liquid phases. To control these equilibria, there are several *extrinsic* and *intrinsic* factors one can alter to obtain the desired effects. Among the extrinsic factor at the physicochemical parameters such as acidity, ionic strength, and solvent composition, i.e. the extrinsic factor focuses on the solvent. The *intrinsic* factors instead focus on the properties of the solute, such as the possession of surface-exposed hydrophilic and hydrophobic functional groups mainly determining the hydrophobicity of the solute. In the case of proteins, the intrinsic factors are particularly related to the amino acid residues found on and near the protein surface and the conformational exchange causing exposure of otherwise buried residues. For the design of protein formulations, altering the solvent conditions seems to be the most applied method to achieve solubility, however it may not always be an option, as some solvent conditions may be either physiologically harmful or incompatible with the experimental methods required for the study of the given protein. Consequently altering the intrinsic factors of proteins may be required in assisting to optimize the solubility of proteins. In specific the usage of site-directed mutagenesis of amino acid residues on the surface of proteins, whose structure has been solved using high-resolution experimental methods such as NMR spectroscopy or X-ray crystallography, can with advantage be applied to substitute poorly hydrated and aggregation-prone residues to well hydrated for something better. Even though the majority of proteins utilize nearly all the naturally occurring amino acids to some extent, it has been shown proteins are highly resistant to mutations given the mutations are semi-conserved i.e. the overall properties of the amino acid residue did not change. Among the prime examples of this, is the restriction of different amino acid residues from 20 to 9 in the 213-residue long protein *Escherichia coli* orotate phosphoribosyltransferase while still maintaining catalytic function.[1]

Based on the previous, the research in this work can roughly be divided up into two categories, i.e. research that focuses on the intrinsic factors and research that focuses on the extrinsic factors. In the following, the discussion of the first three papers presented in this work (papers I- III) will resolve mainly around the intrinsic factor: electrostatics. This is done using two different systems namely the EXG protein family, possessing the unique trait it can be rendered completely changeless, and a Tröger's base-linked bis-crown either, possessing the trait it can bind two like-charged potassium cations. Following the discussion of intrinsic factors and electrostatics, a discussion of the remaining papers (papers IV-VII) will be conducted mainly having an emphasis on the extrinsic factor of adding of co-solvent. This is illustrated in multiple systems including the inorganic molecule cobaltabisdicarbollide, the proteins lysozyme and histatin 5, and finally caffeine.

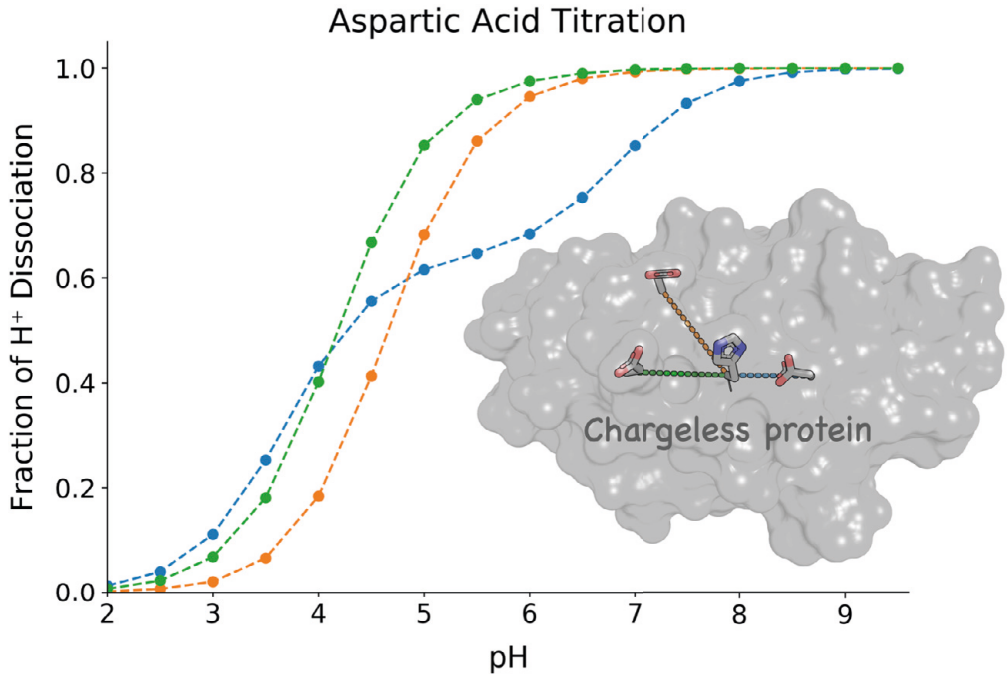



**Figure 5.1:** Titration curves of the individual aspartate residues (39D blue, 43D orange, and 61D green) in the presence of a histidine residue obtained using constant pH molecular dynamics (CpHMD). It is worth noting that while 43D and 61D exhibit single transition Henderson-Hasselbalch titration behavior, 39D exhibits double transition behavior. This observation was also observed using NMR spectroscopy on the systems and motivated the usage of a diprotic acid cycle to investigate the strong linkage between 39D and 66H.

## 5.2 Charge Interactions in a Highly Charge-depleted Protein

Electrostatic forces are important for protein folding and are favored targets of protein engineering due to the shear strength of the interaction and persistence over distance. The persistence over distance is also what causes difficulties when attempting to do rational modifications to proteins, due to the perturbation of a highly complex network of interactions. As of consequence, in paper I we choose to study pairs of titratable residues in a protein otherwise free of such residues (the EXG protein system[47]), to reduce the complexity. Using constant pH molecular dynamics, NMR spectroscopy, and thermodynamic double mutation cycles we were able to give a detailed view into the thermodynamics and structure of the interaction formed between histidine-aspartate pairs.[46] Of the three studied histidine-aspartate pairs, one, in particular, stood out from the rest; the 39D-66H variant (Fig. 5.1). This pair engaged in salt bridging and was found to have a coupling energy of 15 kJ/mol obtained from a diprotic acid cycle. The partition function used to describe individual macrostates within the diprotic acid cycle was constructed from the four possible combinations of protonation states. Choosing the fully deprotonated state as a reference

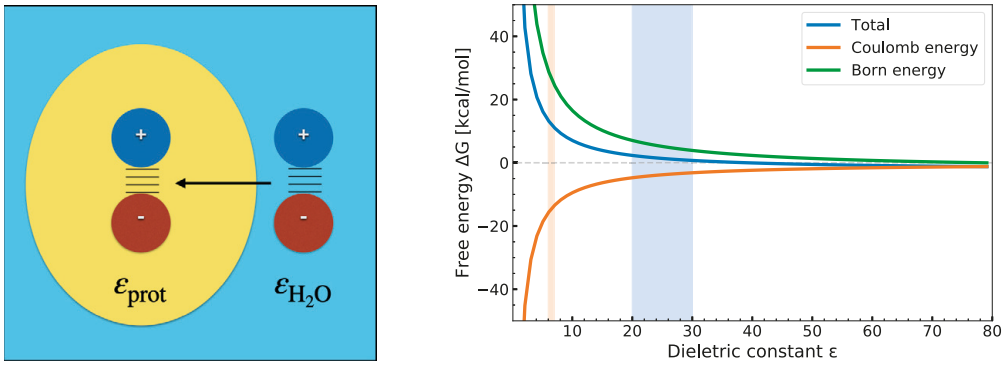



**Figure 5.2:** Simple continuum model constructed from a Coulomb electrostatics term and Born solvation term (see eq. 5.2) of ions for describing the transfer of ion pairs from water ($\epsilon_{H_2O} = 78.4$ at 298 Kelvin) to a lower dielectric environment $\epsilon_r$. Typical values for the dielectric constant of proteins ($\epsilon_{prot}$) are 20-30 near the protein surface at the protein-water interface (blue highlighted area), and 6-7 in buried hydrophobic pockets (orange highlighted area)[66].

state the partition function is given by

$$Z = 1 + e^{-\beta\left(G^0_{01} - \mu_{H^+}\right)} + e^{-\beta\left(G^0_{10} - \mu_{H^+}\right)} + e^{-\beta\left(G^0_{11} - 2\mu_{H^+}\right)}, \tag{5.1}$$

where $G^0$ is the free energy of protonation with the subscript specifying the protonation state vector and $\mu_{H^+}$ is the chemical potential of protons. By fitting the populations of the individual state to the simulation data, a full set of thermodynamic data (free energies) could be extracted including the interaction energy between the residues.

Post-publication reflections include mainly two points. One is the solubility of a salt pair in a lower dielectric, while the second involves the methodology chosen for the study. In regards to the first point, globular proteins are known to be in an equilibrium between a folded and unfolded state. While ion pairs and salt bridges stabilize the folded state, the solvation of ions in a non-polar environment is an unfavorable process, thus stabilizing the unfolded state. In specific we can create a primitive continuum model, constructed from a Coulomb and Born term, which describes the process of transferring ion pairs between different dielectric environments. For proteins we here look at the transfer of an ion pair from water (78.4, at 298 Kelvin) to a lower dielectric

$$\Delta G = \underbrace{\frac{1}{4\pi\epsilon_0\epsilon_r}\frac{q_1 q_2}{r}}_{\text{Coulomb energy}} + \underbrace{\sum_{i=1}^{2}\frac{-q_i^2}{8\pi\epsilon_0 r_{i,0}}\left(\frac{1}{\epsilon_{\text{water}}} - \frac{1}{\epsilon_r}\right)}_{\text{Solvation energy}}. \tag{5.2}$$

Notable it can be seen given the simple model and the parameters chosen it is an unfavorable process transferring ion pairs from a high dielectric to a low dielectric. However,

the free energy penalty of the ion pair being on the surface (blue area highlighted) is much smaller than being buried (orange area highlighted) and thus ion pairs on protein surfaces can be found in stable proteins. Based on this simple model, it proposes two consequences of burying titratable residues in proteins; one consequence is the change in the proton tautomerism equilibrium such that one goes from stabilization by salt-bridging to stabilization by hydrogen bonding. The second consequence is a change in the structural stability of the protein shifting the equilibrium towards to unfolded state due to better solvation of the charged residues in the unfolded state or population of a different structural ensemble characterized by higher dielectric surrounding the newly created ions.

This first mentioned consequence has already been witnessed experimentally by NMR spectroscopy, in which a titratable pair was buried deep within the *staphylococcus nuclease*[106]. In specific it was found the neutral state to be more populated than the zwitterionic state (60/40) for two strongly coupled titratable residues, which is opposite to what was found for the Asp-His pair in the EXG protein. The two systems behavior highly identically, with the titration curves for the titratable residues being highly coupled on a residual resolution, however with the biggest difference between the systems being the degree of burial. The second consequence has also been observed within the *staphylococcus nuclease*.[57] In specific it was shown the propensity for the titratable residues to reorient themself into a more hydrophilic environment, was one of the key factors determining whether or not the structural stability would be affected. For protein engineering, the interplay between dehydration of charges, Coulomb interactions, and the possibility for protein reorganization is thus aspects to be considered in the attempt to design novel, stable proteins.

The second post-publication reflection resolves around the methodology. The constant pH molecular dynamics implementation used had two drawbacks; one being the usage of an implicit solvation model, in specific generalized Born solvation, to calculate the protonation energies instead of maintaining an explicit solvation model, and the second being the method *potentially* incorrectly sampling the Boltzmann distribution. These two drawbacks are in some sense intertwined and both related to the historic development of the discrete CpHMD method. In specific protein simulations utilizing generalized Born solvation for both the sampling of the protonation state space and configurational space was initially utilized due to implicit solvent being computational more feasible than explicit water, however, attempting to sample the protonation state space using explicit water lead to large differences in energy causing the rejection of nearly all proposed protonation state changes. The large energy differences were found to originate from the solvent configuration between protonation and deprotonated state. As of consequence, a hybrid implicit/explicit scheme was proposed, in which the protonation state space was sampled using generalized Born solvation and the configurational space using an explicit solvent. Finally, to address the problem of solvent configuration, solvent relaxation was conducted subsequent to the Monte Carlo move if the protonation state change was accepted to avoid large differences

in energy. However, since the solvent relaxation is not a part of the acceptance criterion, there is a possibility to accept high energy solvent configurations with a higher probability than the Boltzmann distribution dictates at the chosen temperature.

A solution to both problems mentioned; the usage of explicit solvation during the protonation state changes and its avoid high energy solvent configurations, is the utilization of non-equilibrium thermodynamics. In particular, the system is driven out of equilibrium by the protonation state change, and through a finite-time non-equilibrium process a configuration is generated. The acceptance rule, to determine the correct acceptance ratio for the non-equilibrium candidate configuration, is related to the non-equilibrium work rather than the instantaneous energy difference. This approach was first described in detail by Stern[127] with the method later generalized and named "nonequilibrium candidate Monte Carlo moves" by Chodera and coworkers.[89] While no attempts to conduct such simulations on the EXG system have been done, it would be interesting to conduct such simulations to see the effect of explicit water coordination in regards to the solvation of salt bridges.

## 5.3 Systematic Electrostatic Perturbation of a Charge-depleted Protein: Correlation between Protein Solubility and Electrostatics

Having done a complete electrostatic characterization of the titratable EXG variants in paper I and having characterized the charged-depleted variant in the past,[47] the natural continuation would be to address the connection between electrostatic interactions on the solubility of proteins. Utilizing ammonium sulfate precipitation experiments in combination with all-atom molecular dynamics simulations and free energy we set out to explore the relationship between protein electrostatics and protein solubility. Interestingly, we found the solubilities were highly chaotic in terms of behavior, and of the II EXG variants only the 3 single-containing aspartate variants could be explained, and with limited satisfaction. In particular paper II illustrates as it was shown in the past increased solubility correlates with negatively charged surface area,[59] however under the circumstance the structural stability is unchanged. While paper II is limited in terms of conclusions it instead brings the question of how to model and predict the effort of protein electrostatics on the protein solubility. To do a statistical mechanical description of protein solubility and to elude the effect of electrostatics we defined the difference in solvation free energy as

$$\Delta\Delta G_{\mathrm{sol}} = -RT \ln \left( \frac{S_{\mathrm{charged}}}{S_{\mathrm{non-charged}}} \right) \tag{5.3}$$

where $S_{\mathrm{charged}}$ is the protein solubility for any charged variant, and $S_{\mathrm{non-charged}}$ is the protein solubility for the charge-depleted variant, measured by protein precipitation methods with extrapolation to obtain the solubility in water and buffer. Among the factors

expected to contribute to $\Delta\Delta G_{\text{sol}}$ are the one-body contribution and two-body contributions.[1] The one-body contribution is associated with the protein's interactions, and in particular, the solvent-exposed amino acid residues, with the surrounding solvent, in this case, water/buffer. The two-body contribution is associated with the preference of the protein to interact with other protein species with the mechanism usually being hydrophobic interactions combined with the exclusion of water, or electrostatic cross-linking. Both the one-body and two-body contributions depend on the acid-base properties (p$K_a$ values) of the protein and the structural stability. While the acid-base properties determine the formation of charge in the protein necessary for the electrostatic cross-linking, the importance of the structural stability is related to the low solubility of the unfolded state, due to the solvent exposure of the hydrophobic amino acid residues as would otherwise be protected in the folded state.[92] Consequently, are current in the process of investigating the one-body contributions using all-atom molecular dynamics and energy-representation theory of solvation, while the protein-protein (two-body contribution) can be investigated by coarse-grained Monte Carlo simulations, allowing to capture the effects of charge-regulation upon protein-protein interactions. In conclusion, as previously mentioned, the paper is currently limited in terms of conclusions on how to develop strategies to improve the protein solubility by altering the intrinsic factors of proteins, however, the paper does illustrate the inherent difficulty and challenges associated with the study of protein solubility. Furthermore, it also reveals how our empirical approach to protein solubility has room for improvements, such that the EXG protein set and future protein models can be studied, understood, and improved in terms of solubility.

## 5.4 Counter Intuitive Electrostatics upon Metal Ion Coordination: Effects of the Solvent and Conformational Change

Allosteric regulation, i.e. the cooperativity or anti-cooperativity transmitted between ligand binding sites upon binding of ligands, is a central topic within the field of protein chemistry, due to many proteins regulating their ligand-binding capacity by allosteric regulation, usually with the mechanism proposed to be large-scale conformational changes. The best example of allosteric regulation is most likely the protein hemoglobin, responsible for the binding and transport of oxygen in the bloodstream, which has been proposed to have allosteric regulation by possessing an open and closed conformation, having different affinities for oxygen, however with conformational change as a responsible mechanism still being an ongoing research question.[119] In paper III we turned to a much smaller system namely a Tröger's base-linked bis-crown ether, which the two crown ether being 18-crown-6, having an affinity for the coordination of potassium ions. Utilizing isothermal titration calorimetry, we discovered the sequential binding of potassium cations to the Tröger's

[1] These contributions are central to the caffeine solubility discussed in the papers V and VI.

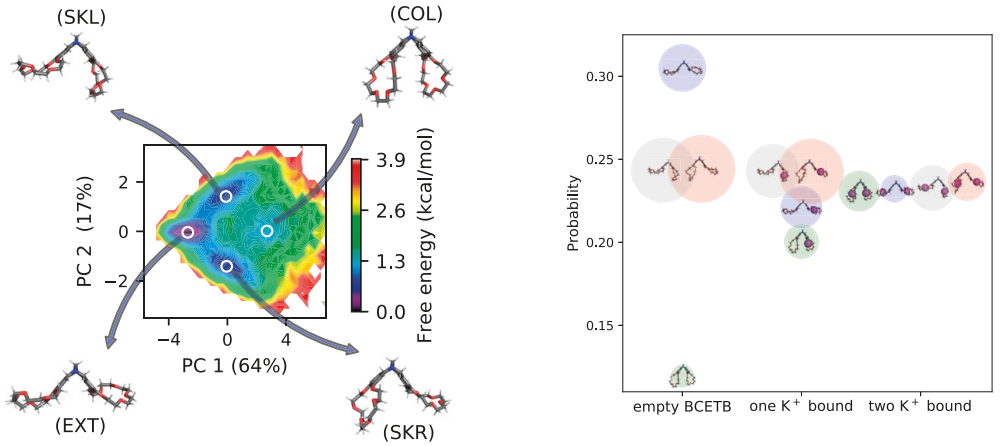



**Figure 5.3:** *Left*: Free energy landscape projected in principal component space, using PCA analysis, of the Tröger's base-linked bis-crown ether with zero bound potassium ions. Notably can be found four distinguishable minima each relating the relative orientation of one of the two crown ethers. It was found 81% of the variance could be explained using the first two principal components with the motion revealed from the first two components physically agreeing with the slowest mode of motion, rendering the usage of more principal components invaluable. *Right*: Perturbation of the relative probability of the individual minima in PCA space upon the binding of one and two potassium ions. The radius of the circles represents the standard error of the mean probability obtained from block analysis.

base-linked bis-crown ether was anti-cooperative $\Delta\Delta G > 0$ being no real surprised due to the cation-cation repulsion, however, decomposing the free energy it was discovered to our surprise that $\Delta\Delta H < 0$ and $T\Delta\Delta S < 0$. For a small molecule most likely incapable of engaging in large scale conformational change, allosteric regulation would seem out of the question, leaving only the regulation by potassium-potassium correlation, however in such case, one would expect $\Delta\Delta H > 0$ due to electrostatic repulsion between the two sites. Consequently, we set out to model the binding potassium to unveil the molecular mechanism associated with the counter-intuitive sign of the thermodynamic state functions.

Using the configurations from the end-state ($\lambda = 0$: uncoupled $K^+$, $\lambda = 1$: fully coupled $K^+$) generated during the free energy calculation, we searched for discrete conformational states of Tröger's base-linked bis-crown ether. Conducting principal component analysis on the non-hydrogen pairwise distance matrix obtained from the equilibrium ensemble we unveiled four discrete conformational states associated with the relative orientation of the crown-ethers (see Fig. 5.3 *left*) in an extended conformation or skewed conformation, with the state of both crown-ethers being in the skewed conformation named the collapsed state. This surprising observation that such a small molecule like the Tröger's base-linked bis-crown ether possess these conformational states ignited the hypothesis that enthalpic cooperativity ($\Delta\Delta H < 0$) of potassium binding could be explained by either the change in equilibrium population of the individual states and/or the change in internal energy for the individual conformational states upon binding of potassium. It was found in particular the collapsed conformational state was the *only* state yielding a negative contribution to $\Delta\Delta H$

and able to overcome remaining positive contributions from the remaining conformational states. The origin of the negative $\Delta\Delta H$ was found to be related to both the change in enthalpy and the shift in the relative population of the individual minima upon binding of potassium. Notably the shift of the rare collapsed conformational state in the fully potassium depleted state to have near-equal probabilities with the remaining states upon the binding of a single potassium cation and equal probabilities upon the binding of the second potassium cation (Fig. 5.3 *right*).

The method chosen computationally to obtain enthalpies is via the van 't Hoff methodology, in contrast to calculating the enthalpy from the end-states only, thus we resolve to conduct simulations at multiple temperatures and utilize the inverse temperature and free energy are linearly related with the enthalpy being related to the proportionality constant given the assumption the change in heat capacity of the system is unchanging upon the binding of potassium. An implication of this is the possibility to utilize temperature replica exchange in parallel with the molecular dynamics sampling to obtain faster converging trajectories and free energy calculations, without added computational cost. However, the utilization of replica exchange, effectively being a Monte Carlo move, disrupts the dynamics of the simulations. This had the ramification that time-dependent structural analysis was no longer possible. In specific the usage of time-lagged independent component analysis (TICA)[49,98,117] over principal component analysis (PCA), which are mainly different from PCA finding principal coordinates of maximal variance while TICA finds coordinates of maximal auto-correlation at a given lag time, with TICA usually being superior to identify the slowest mode of motion.[90,98] This had the consequence that the individual structural states in TICA space are much more displaced from one another compared to in the PCA space, making the integration over the reduced space a trivial task to obtain the probabilities of the structural states. Consequently, it would be interesting to develop and explore possible schemes in which one can combine enhanced sampling with the powerful dimensionality reduction method TICA.

Up to this point, the explanation of the counter intuitive electrostatics and thermodynamic state functions have been explained by the internal interactions within the Tröger's base-linked bis-crownether, however from continuum electrostatics another contribution is possible. The Coulomb potential previously presented as the first term of Eq. 5.2 relates the reversible work of bringing two charges infinitely far from each other, being the ground state, to a given certain separation $r$. For like-like charges, the change enthalpy is given by

$$\Delta H_{++} = \frac{\partial}{\partial T^{-1}} \left( \frac{\Delta G_{++}}{T} \right) = \Delta G_{++} - T \frac{\partial \Delta G_{++}}{\partial T} \tag{5.4}$$

The Coulomb potential integrates the degrees of freedom from water out represented by the dielectric constant, $\epsilon_r$, being a value representative of the liquid's tendency to orientate it self to oppose an external electric field, and is thus temperature dependent. Inserting the

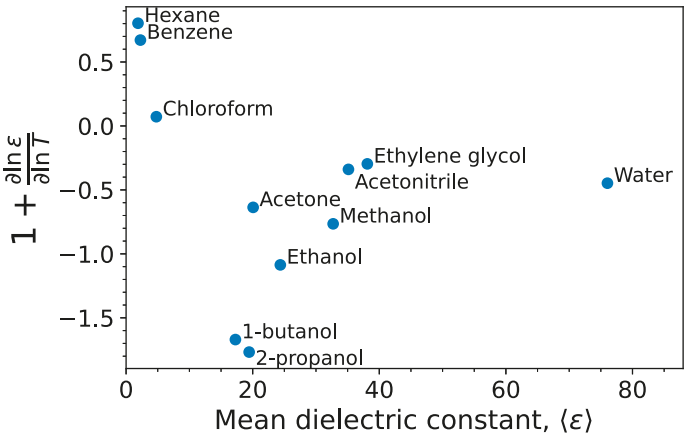



**Figure 5.4:** Scatter plot of the logarithmic variation in dielectric constant with respect to the logarithmic temperature in Kelvin and the mean dielectric constant, $\langle \epsilon \rangle$, over the studied temperature range for a selection of various simple solvents. The logarithmic variation in dielectric constant with respect to the logarithmic mean temperature enters into the expression for the enthalpy and entropy decomposition of the free energy calculated from the Coulomb potential for ions of like charge (see Eq. 5.5 for enthalpy expression). In particular when the derivative plus one is negative $\Delta H$ will also be negative, while positive values cause $\Delta H$ to be positive.

first term of Eq. 5.2 and factoring out everything but the dielectric constant, one arrive at

$$\Delta H_{++} = \Delta G_{++} \left( 1 + \frac{\partial \ln \epsilon_r}{\partial \ln T} \right) \tag{5.5}$$

The partial derivative of the logarithmic dielectric constant with respect to the logarithmic temperature may be evaluated experimentally and is only dependent on the choice of solvent. Specifically for water, the value is found to be -1.36, and thus the continuum model correctly captures the sign of the state functions equivalently to what was found experimentally. The explanation to this has been provided earlier by Israelachvili,[50] proposing the increase in entropy and decrease in enthalpy shall be understood from the ordering of the water around the like-charged ions in the attempt to counteract the unfavorable electrostatic interaction. However, in such a case it would be interesting to know which solvents possess the capacity to counteract electrostatics in such a manner. Eq. 5.5 proposes the simple answer is simply to monitor liquids dielectric response with respective with temperature. An a selection of various simple solvents has been visualized in Fig. 5.4.

## 5.5 Total Description of Intrinsic Amphiphile Aggregation: Calorimetry Study and Molecular Probing

The aggregation and creation of micelles by amphipathic solute is a well-known process with perhaps the most well-known example being the formation of soap. The thermodynamic

driving force of micellization is associated with the hydrophobic effect which is commonly attributed to being entropic of nature however the molecular aspects of the hydrophobic effect are not fully understood. Amphipathic solutes are characterized by the molecule having a dual-polarity, that is one part of the molecule is hydrophobic with another part being hydrophilic. Commonly, amphipathic molecules which engage in micellization, are said to possess a ”head and tail” design in which the ”head” constitutes the hydrophilic region while the ”tail” constitutes the hydrophobic region. Thus it has been proposed the ”head and tail” structure of amphipathic molecules is a prerequisite for the formation of micelles. However, it has been experimentally found that cobaltabisdicarbollide (COSAN) anions possess a high affinity for self-assembly, creating vesicles and micelles, despite the molecular architecture of COSAN not having a somewhat obvious ”head and tail”-design. As of consequence, the structural and thermodynamic properties of COSAN micellization have been excessively investigated in the last decade. In particular, it has been found experimentally that COSAN micellization is an enthalpy-driven process,[39] whereas micellization of surfactants is commonly associated with being an entropy-driven process.[34]

In paper IV we employed a range of experimental and computational techniques to fully understand the self-assembly process of COSAN, while also addressing the effect of altering the solvent conditions by the addition of acetonitrile. Using NMR spectroscopy and all-atomic molecular dynamics, we uncovered that acetonitrile enhances the aggregation of COSAN through dipole-dipole interactions and by altering the solvation shell, while not co-aggregating with COSAN.

While experiments like NMR spectroscopy provides some insight into the molecular-scale understanding of COSAN aggregation in aqueous solution, the contribution of molecular simulations is essential to elude the molecular mechanism of this non-classical (enthalpy driven) hydrophobic effect. Disagreement however arises in regards to the structural properties of COSAN from molecular simulations, in particular as to what constitutes the hydrophobic and hydrophilic regions of COSAN. In specific our model finds boron-bound hydrogens to be negative, as opposed to partially positively charged carbon-bound hydrogens in organic compounds.[136] However in a recent study the exact opposite was observed; that boron-bound hydrogens are positive and boron atoms negative.[70] Thus, one post-publication reflection is the discrepancy in regards to the parameterization of COSAN for molecular simulation.

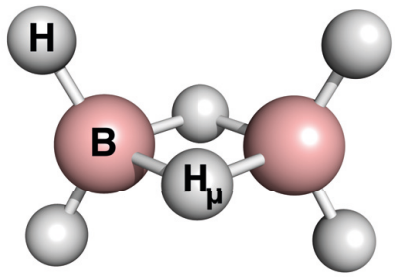
In particular, for the classical atomistic modeling of electrostatic interactions, the majority of force fields utilize point charges which are associated with the coordinates of individual atomic nuclei. While the Schrödinger equation provides the quantum mechanical description of molecules as positive nuclei surrounded by a negatively charged electron cloud, traditional chemistry, and classical simulation methods are still operating with the concept of atoms joined by chemical bonds with atoms having a net charge to describe the polarity of the chemical bond. This quantum to classical mechanical approximation of charge dis-

tributions associated with atomic nuclei is referred to as *atomic charges* or *partial charges*. The idea of effective atomic partial charges in molecules and crystals is very appealing to describe ionicity or polarity of chemical bonds and molecules and is widely used though-out different fields of chemistry, however, due to partial charges being a simplification of reality no true answer exists as to what the partial charges of atoms are and thus multiple schemes exists to calculate partial charges. The two most common methods for the determination of partial charges, for atomistic molecular simulation, is population analysis of wave functions based on basis functions and population analysis based on electrostatic potentials. Both of these are characterized by utilizing *ab initio* calculations in which the choice of level of theory and basis set will influence the result of the calculation. Among the methods relying on population analysis of wave functions, the choice of basis set and level of theory is most dramatic for Mulliken population analysis,[85] which does not converge with increasing basis set,[72,142] as electronic wave functions far from the nuclei of interest, would still be counted as belonging to that nucleus instead of other surrounding nuclei.[52] However alternative schemes utilizing population analysis of wave functions have addressed this issue including partial charges obtained using the Natural Population Analysis (NPA)[103] method. Yet, despite the NPA method being an excellent choice for analysis of chemical bonds due to its rigorous quantum mechanical framework, one major shortcoming is the overprediction of polarity in hydrogen-carbon bonds,[52,142–144] which is commonly considered to be of covalent bond character. This problem was addressed by Reed & Weinhold, that the individual bond dipole moments would potentially not add up to the total dipole moment due to improper treatment of the distribution of charge related to the different atomic orbitals and quality of charge related to the polarization of orbitals with all issues arising from approximating charge distributions with fixed points.[102] While some all-atomic force fields utilize population analysis of wave functions like the CHARMM force field,[69] other force fields utilize population analysis based on the electrostatic potential (ESP) with one example being the AMBER force field.[26] One of the major drawbacks in the utilization of charges derived from fitting charges to electrostatic potential is the fitting procedure itself. In particular, it can cause a challenge to find the best fit possible to recreate the electrostatic potential. As a consequence, it is not uncommon that equivalent atoms, due to symmetry possess different charges, which has no physical meaning, but an artifact arising from the fitting of charges.

To show the impact of choice of population analysis, in specific NPA versus ESP derived charges, a set of partial charges has been determined for the molecule diborane, which is a much smaller molecule with comparable hydrogen-boron bonding to COSAN (Tbl. 5.1). It is worth noting that while the charge for $H_\mu$ is very much equivalent for the two population analysis methods despite the basis chosen, the main differences for the methods occur for the boron (B) and single boron-bound hydrogen atoms (H). Just as with COSAN, we find the charge of boron and single boron-bound hydrogen have been inverted for all correlation-consistent basis set higher than double zeta, and double and triple zeta Pople ba-

**Table 5.1:** Partial charges for the boron and hydrogen atoms in diborane ($B_2H_6$). The geometry was optimized using B3LYP/6-31G* with the partial charges evaluated using B3LYP and the listed basis set and population analysis in the table. Equivalent atoms due to symmetry have been averaged with the atomic naming given in the illustration below the table

| | B | | $H_\mu$ | | H | |
|---|---|---|---|---|---|---|
| Basis set | NPA | ESP | NPA | ESP | NPA | ESP |
| 6-31G | -0.187 | 0.0143 | 0.14 | 0.121 | 0.0235 | -0.0678 |
| 6-311G | -0.101 | 0.0605 | 0.0954 | 0.104 | 0.00268 | -0.0821 |
| cc-pVDZ | -0.0908 | -0.0154 | 0.1 | 0.146 | -0.00472 | -0.0651 |
| cc-pVTZ | -0.121 | 0.0515 | 0.108 | 0.117 | 0.0065 | -0.0843 |
| cc-pVQZ | -0.114 | 0.0496 | 0.105 | 0.117 | 0.00452 | -0.0835 |
| aug-cc-pVDZ | -0.0774 | 0.0942 | 0.0915 | 0.0966 | -0.00703 | -0.0954 |
| aug-cc-pVTZ | -0.12 | 0.0521 | 0.107 | 0.116 | 0.00618 | -0.0839 |
| aug-cc-pVQZ | -0.116 | 0.0512 | 0.106 | 0.116 | 0.00493 | -0.0836 |




sis set. As previously stated, due to the concept of partial charges being a fictitious one, both the sets of atomic charges may be reasonable solely based on the information available from *ab initio* quantum mechanical calculations for the modeling of COSAN. Consequently, to gain insight into the correct charge distribution, experimental insight is required. One option is to investigate the preferential orientation of a dipole molecular in the presence of COSAN. From the molecular dynamics simulations using charges obtained from electrostatic potential fitting, it was found acetonitrile is preferentially aligned to COSAN, with the methyl groups being the main site of interaction with the boron atoms of COSAN, based on radial distribution function peaks correlating with the B-B atom distance. This observation was utilized to explain the NMR data, in specific the chemical shifts obtainable from $^{1}$H, $^{13}$C, and $^{1}$H–$^{15}$N heteronuclear multiple bond correlation NMR, which seemed to suggest the nitrile group to possess more rotational freedom. This preferential alignment is most likely attributed to the choice of parameterizing boron as positively charged spheres, however, it remains unknown if we could obtain almost similar conclusions using a force field utilizing NPA-derived charges. Two possible hypotheses and questions are essential to address this question; I) Given the NPA charges, would we find acetonitrile to preferentially associate itself to the negatively charged boron atoms or the positively charged hydrogen atoms? II) What is the preferential orientation of acetonitrile?

While molecular dynamics in our case has been utilized to explain experimental data and molecular phenomena, one common and essential requirement is to establish the valid-

ity of the model utilized in the reproduction of experimental data by the simulation, such that molecular mechanisms can be derived and interpolation and extrapolation can be conducted with confidence. While no such efforts were rigorously conducted using the ESP-derived charges for COSAN, the osmotic pressure was reproduced with charges derived from NPA, thus making the study rest on a more solid foundation. Consequently, it would be interesting to attempt to reproduce the osmotic pressure given the ESP-derived charges for COSAN. This effort combined with the previously addressed questions would greatly assist in constructing a rigorous model for COSAN for future prediction of COSAN's interaction with molecular matter.

## 5.6 Statistical Thermodynamic Description of the Molecular Solvation of Caffeine in Salt Solutions

Caffeine is presumably the most consumed psychoactive drug worldwide,[99] most commonly found in coffee, tea, and energy drinks. Despite caffeine typically being characterized as a bitter taste stimulant,[100] the caffeine-containing beverages are paradoxically considered by many a great joy. Therefore the process of coffee brewing has undergone tremendous development and experimentation to obtain correct amounts of caffeine by extraction from solid to the aqueous phase to create the perfect cup of coffee.[25,80,140] Consequently, it is desirable to understand the physical and chemical proprieties of caffeine to optimize and understand processes in which caffeine is involved, including drink brewing such as coffee and tea, medicine, and other industrial, pharmaceutical, and biological proposes. Caffeine is possessing a chemical structure highly related to the purine nucleobases of DNA and RNA, the physical properties of caffeine have been exhaustively investigated by a great variety of methodologies, including experimental, computational, and theoretical methods. In particular, it is known that caffeine is surprisingly soluble in both polar solvents with a preference for chloroform over water,[118] while only sparsely soluble in non-polar organic solvents, due to caffeine's molecular structure being highly heterogeneous in terms of polarity. Additionally, caffeine has also been found to possess a self-association equilibrium, forming highly ordered oligomers characterized by the face-to-face stacking of the xanthine motif of caffeine,[113,114,132,133] highly equivalent to the stacking of the nitrogenous bases found in DNA and RNA.[141] However the formation of larger aggregates has also been reported, in which the oligomers are also branched at the methyl groups.[134] The mentioned equilibria; the partitioning of caffeine in the aqueous phase and organic phase, and the self-association of caffeine are all subject to modulation by osmolytes, such as sugars,[67,120,122] and salts.[54,108,120]

In paper V we investigated the molecular mechanism underlying the solvation of a caffeine monomer in water and salt solutions using molecular dynamics and energy-representation

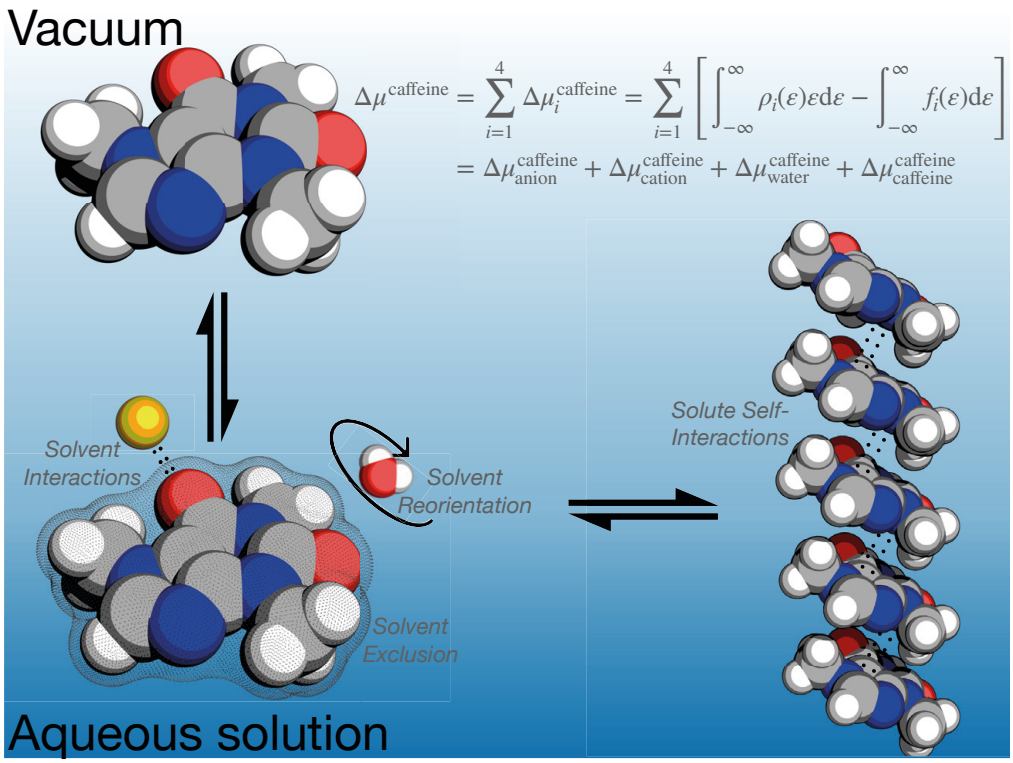



**Figure 5.5:** Main concepts of the focus of study in papers V and VI. In particular, paper V focuses on the vertical equilibrium, in which a single caffeine monomer is transferred from a vacuum to a caffeine-free aqueous solution containing a finite amount of salt. From the mathematical expression in the upper-right corner, each species constituting the solvent, i.e. anions, cations, and water each contribute to the chemical potential of caffeine, which the mechanisms of actions partitioned into the contribution from solvent interactions, work originating from solvent reorientation, and work from solvent exclusion by caffeine. Finally, upon transferring caffeine from a vacuum to a non-caffeine-free aqueous solution, the preexisting caffeine molecules also constitute the solvent, thus yielding the extra contribution of solute self-interaction to the chemical potential of caffeine. This self-association equilibrium, as visualized by the horizontal equilibrium, is the main focus of paper VI.

theory of solvation.[76–78] Due to the usage of the energy-representation theory of solvation framework, the chemical potential of caffeine may be written as

$$\mu^{\text{caffeine}} = \sum_{i=1}^{3} \mu_i^{\text{caffeine}} = \sum_{i=1}^{3} \left[ \int \rho_i(\varepsilon)\varepsilon \mathrm{d}\varepsilon - \int f_i(\varepsilon)\mathrm{d}\varepsilon \right], \tag{5.6}$$

where $\rho(\varepsilon)$ is the probability density of the given pair-energy $\varepsilon$ and $f$ is a function of the pair-energy and is related to the work associated with solvent reorganization and reorientation, covered in more detail in chapter 3.3.4. The summation in Eq. 5.6 is taken over the number of solvent species in the solvent, which in the case of caffeine in salt solutions constitutes three species; water, anions, and cations. Eq. 5.6 is the foundation for two decompositions of the chemical potential. One is the decomposition strategies is by

decomposing the summation over the individual species (specie-decomposition) thus allowing to estimate the effect of the individual solvent species to the chemical potential of caffeine. Another decomposition strategy is to decompose the integration over the energy coordinate into multiple terms (energetic-decomposition), thus allowing the estimation of the contribution to the chemical potential of caffeine from the various energetic regime. The most obvious energetic divisions would be at the highest pair-energy between caffeine and solvent observed from a molecular simulation, thus meaning the probability distribution of pair-energies taken at this value or higher has the probability zero, thus rendering the first integration term zero from the maximum pair-energy to infinity. Using the outlined decomposition strategies it was discovered using monovalent salts from various parts of the Hofmeister series that anions fundamentally increases the chemical potential of caffeine thus increasing the solubility of caffeine in the aqueous solution (i.e. *salting-out*) while cations fundamentally decrease the chemical potential thus decreasing the solubility of caffeine in the aqueous solution (i.e *salting-in*), with the effect of the ions following the well-known direct Hofmeister series (Cations: $Na^+ < K^+ < Cs^+$; Anions: $I < Cl^- < F^-$). The mechanism of action by the individual species was unveiled using the energetic-decomposition strategy and structural properties like radial distribution to discover the mechanism of the cations was found to be associated with the binding of cations to the polar ketone groups of caffeine, while the anions were found to be associated with the modulation of water, due to anions stronger electrostatic interactions with the hydrogen of water, compared to the electrostatic interactions between cations and the oxygen of water. The effect of anions is in great agreement with previous findings in the literature, that is the perturbation of water structure by anions, however interestingly the cationic effect is commonly omitted and/or considered non-existent in experimental studies. By correlating the variation in the chemical potential with the salt concentration for the various salts with the contribution to the chemical potential arising from excluded volume and interactions, it was found both correlations yielded points clustered around the same anion, suggesting the whole perturbation of the solubility of caffeine by salt is governed by the anion. Due to the inseparable contribution of anions and cations by experiments this conclusion is in great agreement with experimental literature and explains the attribution of ion-specific effects predominantly to anions.[54,108]

While paper v puts focus on the solubility of caffeine in the view of transferring a caffeine monomer from a vacuum to an aqueous solution with salt increasing or decreasing the preference for the caffeine monomer to remain in the aqueous solution, paper vi instead puts focus on another solvation process: the aggregation of caffeine. It was previously mentioned caffeine forms highly ordered smaller aggregates, characterized by face-to-face stacking of the caffeine monomers, at the formation of even larger aggregates branching occurs at the hydrophobic methyl groups of caffeine. While this process has been studied in the past, we desired to understand how salt can modulate this equilibrium. Using atomistic molecular dynamics, coarse-grained Metropolis Monte Carlo simulations, and

vapor pressure osmometry we managed to gain insight into the structural properties of the liquid; in specific the formation of aggregates and to reproduce thermodynamic experimental data such that the simulations can be utilized to derive mechanistic insight. By combining the excess chemical potential and its decomposition into one-body and two body from Monte Carlo simulations, with vapor pressure osmometry we show using Kirkwood-Buff inversion, the self-association of caffeine is diminishing effect of salt, thus causing the Setschenow coefficient to become caffeine-concentration dependent, explaining the differences in Setschenow coefficient observed for experimental methods dealing with caffeine solubility at dilution or saturation.

## 5.7 Stabilization and Aggregation of Proteins by Poly Phosphate-Compounds

It is well known that proteins' structural stability and aggregation propensity depends strongly on the solvent and co-solvents such as urea, guanidinium chloride, and ammonium sulfate (see chapter 2.2.1). In the last couple of decades, re-entrant liquid condensation of proteins has been investigated with various kinds of co-solvents. In particular, it was found that salts possessing highly charged ions such as the trivalent yttrium ion ($Y_3^+$) can cause the aggregation of the protein human serum albumin (HSA).[55,146,151] The anomalous phenomenon however is, that despite moderate concentrations of the salt yttrium chloride ($YCl_3$) caused aggregation, increasing the salt concentration even further caused HSA to reenter the liquid phase. The mechanism of this phenomenon has been attributed to the idea proteins in the absence of salt are repealing one another by the charged residues on the protein surface yielding the double layer, with the attraction at moderate concentrations of highly charged solvent is governed by a balance of interactions between charged co-solvent ions and charged residues on the protein surface to essentially cross-link proteins, while at even higher concentrations of the highly charged co-solvent the protein surface is saturated with ions causing a charge inversion thus causing the proteins ones-more to repel one another.[95] The attraction between proteins at moderate concentrations is a mixture of electrostatic interactions and hydrophobic interactions with the ratio between the two modulated by the addition of monovalent salt, which has no strong preferential binding to protein sites and thus only having the role of increasing the electrostatic screening.[58,95] Given this knowledge it would seem the charge of the co-solvent to be the only parameter of importance in causing charge inversion, however by Bye and coworkers it was revealed that only preferential polyvalent salts could cause the reentered condensation of hen egg-white lysozyme (HEWL).[20] Among the salts to have the property of causing reentered condensation were sodium diphosphate and sodium triphosphate, while the salt sodium citrate was unable to cause reentered condensation. A possible explanation for the causing of re-entrant liquid condensation of HEWL, is due to the intrinsic affinity of arginine to

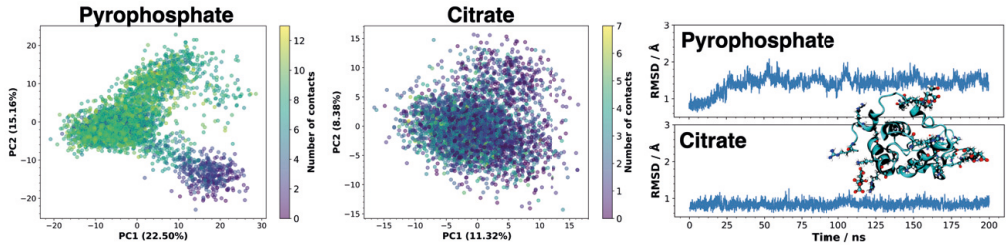



**Figure 5.6:** Structural analysis of lysozyme in the presence of pyrophosphate/diphosphate (-4*e*) and citrate (-3*e*) of the protein dynamics obtained from atomistic molecular dynamics simulations. Principal component analysis utilizing non-hydrogen backbone atom pairwise-distances of lysozyme in the presence of (*left*) pyrophosphate and (*middle*) citrate with the color map yielding the number of bound anions within 6 Ångstrom of the side-chain functional group of lysine or arginine. (*Right*) Root mean square deviation of non-hydrogen backbone atom positions of lysozyme over time, with the reference structure being an equilibrated structure of lysozyme in the absence of citrate and pyrophosphate.

phosphate, as arginine-rich motifs are commonly found in nucleotide-binding proteins and kinases. [148,149]

To investigate whether the importance of a protein surface residues' affinity to the co-solute was of importance we utilized the antimicrobial, salivary intrinsically disordered protein Histatin 5 (Hst5), [101] which naively possesses three arginine residues. The utilization of this protein allows two theories to be explored: (I) whether it is possible to cause re-entrant liquid condensation of intrinsically discorded proteins and (II) whether strong affinity between co-solute and protein surface residues are of great importance. It was found it is indeed possible to cause re-entrant liquid condensation of Hst5, and by arginine-to-lysine substitutions, it was found Hst5 was much more prone to aggregation in the presence of arginine over lysine. [65]

Up to the initiation of the project, we had previously already engaged in unpublished research regarding the high affinity for pyrophosphate to arginine amino acid residues. In particular, we investigated the association and structural response of lysozyme in the presence of the deprotonated state of pyrophosphate, having the charge -4*e*, and citrate, having the charge -3*e*. It was found that citrate and pyrophosphate both were found to associate themself to the positive residues of lysozyme being arginine, lysine, and histidine, as somewhat expected. Interestingly, however, we found the association of pyrophosphate to lysozyme induced a structural perturbation, as it is observable from principal component analysis (PCA) based on non-hydrogen protein backbone atom pairwise distances and root mean square deviations (RMSD) of the protein backbone atoms. In figure 5.6 one can see lysozyme throughout 200 nanoseconds is adapting a configuration, which possesses different structural characteristics than those found in the crystal structure when in the presence of pyrophosphate. These events are evident from the two distinguishable "islands" in the PCA space for lysozyme in the presence of pyrophosphate, as well as a slow transition in structure for approximately 30 nanoseconds when viewing the RMSD as a

function of time. In contrast, we found that citrate did not induce any discrete conformational change, but the dynamics are mainly associated with fluctuations in the structure from the crystal structure, due to the finding of a single minimum in PCA space and no obvious transitions between states from the RMSD versus time. To ensure the transition in the structure of lysozyme by pyrophosphate was due to binding, and not electrostatic screening,[5] weakened hydration,[107] or other bulk effects related to the hydration of salt in aqueous solution, the number of contacts between pyrophosphate and the cationic residues of lysozyme was tracked as a function of time. Overlaying the number of contacts between pyrophosphate and citrate to the cationic residues of lysozyme as found from the various frames in the trajectories, with the frames in PCA space. It is observed the structural change is indeed corresponding to the binding of pyrophosphate as the region associated with the crystal structure and the generated region are characterized by having a lower and higher number of contacts respectively. Visualizing the trajectory, it was evident that arginine was the main site of attraction, compared to that of lysine, which could potentially be attributed to arginine possessing a guanidino group in the side chain, providing increased salt bridging capacity, compare to lysine's ammonium group. Even though no efforts have been done to attempt the validity of the force forces utilized, and the transition between the states may not appear reversible on the timescales studied here and is thus unsuitable for an equilibrium statistical thermodynamics treatment, we anticipated the findings to be of great interest and importance, culminating with the hypothesis of strong arginine affinity to pyrophosphate being a substantial contribution to the aggregation and reentering condensation of Hst5.

## 5.8 Contextualization and Future of Solvation Thermodynamics

At the start of this chapter, the quote "*Knowledge is knowing a tomato is a fruit; wisdom is not putting in a fruit salad*" was presented. While we have in the former discussed the research in terms of mechanisms, properties, and so on, all of which expand our knowledge, to achieve wisdom we need to be able to apply our knowledge into context. Therefore, here we will conduct a contextualization of the theory previously presented (chapter 1-4) and the research conducted in this work (chapter 5 and papers).

Having a background in experimental protein chemistry before transitioning to theoretical and computational protein chemistry, yields the possibility of viewing problems from both the perspective of an experimentalist and theoretician. In particular, the individual disciples are well aware of their limitations in terms of obtainable knowledge regarding chemical systems, and hence collaboration is in the majority of cases beneficial to investigate problems. The majority of experimental groups tend to adapt and investigate properties of proteins that are considered "well-behaving", meaning the proteins are expressed in good amounts *in vivo* and do not form aggregates or engage in a conformational change in optimal exper-

imental buffer and salt conditions. Yet an equally interesting field of study is the turning of ”ill-behaved” proteins into ”well-behaved” proteins by rational modifications of proteins or solvent conditions. As an example, with the investigation of polyphosphates' effect on Hst5 and lysozyme, we can now rationalize the formation of Hst5 dimers when solvated in phosphate buffers. With the knowledge that phosphate interacts strongly with surface-exposed arginine residues, one has the option of doing arginine to lysine substitutions if possible, or alter the choice of the buffer. The methods utilized in this work, namely molecular dynamics, Monte Carlo simulations, and energy-representation theory of solvation, yield a strong toolkit attempting to predict and understand how to turn ”ill-behaving” proteins into ”well-behaving proteins” on an atomic resolution by varying the many intrinsic and extrinsic factors that exist. Consequently, I urge the experimental protein chemistry community to not disregard data of ill-behaving proteins, but instead investigate what causes them to exhibit these properties in collaboration with theoreticians. While the current trend today is still ”negative data” are harder to publish than ”positive data”, efforts to break this cycle is being attempted, and at some point, these studies need to be addressed, as they can have a major impact on industrially used proteins and design of new novel proteins. At the same time, computational studies of proteins are usually conducted at the minimal conditions of co-solute. For example, buffer molecules are usually omitted in many computational studies, due to the assumption the buffer does not interact with the protein of interest, however as we saw from lysozyme and COSAN, small concentrations of additive can cause dramatic changes to equilibria. Therefore, it is also recommended to consider the design of computational experiments, in the attempt to elude factors of importance for the given equilibrium that is studied. For example, the effect of ionic strength can be studied at the level of explicitly including electrolyte species, but may also be studied at the level of the continuum by Debye-Hückel theory, with the difference between the two yielding insight into non-electrostatic effects attributed to the specific electrolyte.

To address the future of the topic solvation thermodynamics, it is worth first looking at the field's previous accomplishments. The first breakthrough, which has also been a continuous theme in this work, is the solvation of electrolytes in water, which is mainly credited first to the work by Peter Debye and Erich Hückel with their development of Debye-Hückel theory, and later on, detailedly explored by Arrhenius and van 't Hoff. Due to the success of working out why salts dissolve in water, the field temporarily succumbs due to its success. The field has now once more reemerged, however, this time facing the challenge of explaining the hydrophobic effect, which has been recognized as the main driving force in the folding of proteins[137], and the concept of hydrophobicity itself. Among the people standing up to the challenge was David Chandler[3] who described the main driving forces of hydrophobic self-assembly[22] but also developed a new theory of hydrophobicity based on the hydration of small and large solutes.[21] Given these success stories of the field, it is expected the field of solvation thermodynamics will continue to flourish. Especially the field should now enter an era rich in the development and application of methods to ex-

plore, understand, and promote new systems beneficial to the industry for the stabilization of new (protein) drugs, chemical processes, and so on. Another aspect, which is a stereotypical opinion of protein chemists,[2] is to address the "specialty" of specific proteins. In particular, many proteins are presented as unique and possessing properties that make them stand out, and thus unable to be compared to other protein systems. As such it would be interesting to study multiple proteins to obtain non-system-specific properties, while simultaneously also attempt to address system-specific properties which may be shared with or transferred to other systems. In light of the previous statement, the PIPPI Ph.D. consortium is most definitely a step in the right direction, with the investigation of multiple proteins under varying physiological parameters such as pH, buffer, ionic strength, co-solute, and so on. However to assist in gaining insight into the protein-protein interactions and protein-excipient interactions and their biophysical consequence computational methods can be of great aid. While interactions between proteins and excipients have been studied at $\lambda = 1$, it could be interesting to add the free energies of solvation to the knowledge pool by the energy-representation theory of solvation. This statement is based on the results shown in Paper v, that more significant contributions may be present in the solvation free energy than the ones one can find from simply conducting simulations at $\lambda = 1$. In conclusion: the future of solvation thermodynamics does indeed look promising, with still many unanswered questions to be answered and many processes which can be optimized for the better of humankind.

[2] Including myself.

# Chapter 6

# References


[1] S. Akanuma, T. Kigawa, and S. Yokoyama. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proceedings of the National Academy of Sciences*, 99(21):13549–13553, October 2002. doi: 10.1073/pnas.222243999. URL `https://doi.org/10.1073/pnas.222243999`.

[2] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, February 1980. doi: 10.1063/1.439486. URL `https://doi.org/10.1063/1.439486`.

[3] Hans C. Andersen, Arup K. Chakraborty, and John D. Weeks. David chandler. 15 october 1944—18 april 2017. *Biographical Memoirs of Fellows of the Royal Society*, 68: 87–102, March 2020. doi: 10.1098/rsbm.2019.0046. URL `https://doi.org/10.1098/rsbm.2019.0046`.

[4] Johan Åqvist, Petra Wennerström, Martin Nervall, Sinisa Bjelic, and Bjørn O. Brandsdal. Molecular dynamics simulations of water and biomolecules with a monte carlo constant pressure algorithm. *Chemical Physics Letters*, 384(4-6):288–294, January 2004. doi: 10.1016/j.cplett.2003.12.039. URL `https://doi.org/10.1016/j.cplett.2003.12.039`.

[5] Sujit Basak, R. Paul Nobrega, Davide Tavella, Laura M. Deveau, Nobuyasu Koga, Rie Tatsumi-Koga, David Baker, Francesca Massi, and C. Robert Matthews. Networks of electrostatic and hydrophobic interactions modulate the complex folding free energy surface of a designed $\beta\alpha$ protein. *Proceedings of the National Academy of Sciences*, 116 (14):6806–6811, March 2019. doi: 10.1073/pnas.1818744116. URL `https://doi.org/10.1073/pnas.1818744116`.

[6] Donald Bashford and David A. Case. Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51(1):129–152, October 2000. doi: 10.1146/annurev.physchem.51.1.129. URL https://doi.org/10.1146/annurev.physchem.51.1.129.

[7] Arieh Ben-Naim. Is entropy associated with time's arrow?, 2017.

[8] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, October 1976. doi: 10.1016/0021-9991(76)90078-4. URL https://doi.org/10.1016/0021-9991(76)90078-4.

[9] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, October 1984. doi: 10.1063/1.448118. URL https://doi.org/10.1063/1.448118.

[10] P. Beroza, D. R. Fredkin, M. Y. Okamura, and G. Feher. Protonation of interacting residues in a protein by a monte carlo method: application to lysozyme and the photosynthetic reaction center of rhodobacter sphaeroides. *Proceedings of the National Academy of Sciences*, 88(13):5804–5808, July 1991. doi: 10.1073/pnas.88.13.5804. URL https://doi.org/10.1073/pnas.88.13.5804.

[11] Thomas C. Beutler, Alan E. Mark, René C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, June 1994. doi: 10.1016/0009-2614(94)00397-1. URL https://doi.org/10.1016/0009-2614(94)00397-1.

[12] Georgios C. Boulougouris. Calculation of the chemical potential beyond the first-order free-energy perturbation: From deletion to reinsertion†. *Journal of Chemical & Engineering Data*, 55(10):4140–4146, October 2010. doi: 10.1021/je100015v. URL https://doi.org/10.1021/je100015v.

[13] Efrem Braun, Seyed Mohamad Moosavi, and Berend Smit. Anomalous effects of velocity rescaling algorithms: The flying ice cube effect revisited. *Journal of Chemical Theory and Computation*, 14(10):5262–5272, August 2018. doi: 10.1021/acs.jctc.8b00446. URL https://doi.org/10.1021/acs.jctc.8b00446.

[14] Efrem Braun, Justin Gilmer, Heather B. Mayes, David L. Mobley, Jacob I. Monroe, Samarjeet Prasad, and Daniel M. Zuckerman. Best practices for foundations in molecular simulations [article v1.0]. *Living Journal of Computational Molecular Science*, 1(1), 2019. doi: 10.33011/livecoms.1.1.5957. URL https://doi.org/10.33011/livecoms.1.1.5957.

[15] Robert Brown. XXVII. a brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21):161–173, September 1828. doi: 10.1080/14786442808674769. URL https://doi.org/10.1080/14786442808674769.

[16] Roland Bürgi, Peter A. Kollman, and Wilfred F. van Gunsteren. Simulating proteins at constant pH: An approach combining molecular dynamics and monte carlo simulation. *Proteins: Structure, Function, and Bioinformatics*, 47(4):469–480, April 2002. doi: 10.1002/prot.10046. URL https://doi.org/10.1002/prot.10046.

[17] Giovanni Bussi and Michele Parrinello. Stochastic thermostats: comparison of local and global schemes. *Computer Physics Communications*, 179(1-3):26–29, July 2008. doi: 10.1016/j.cpc.2008.01.006. URL https://doi.org/10.1016/j.cpc.2008.01.006.

[18] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007. doi: 10.1063/1.2408420. URL https://doi.org/10.1063/1.2408420.

[19] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. *The Journal of Chemical Physics*, 130(7):074101, February 2009. doi: 10.1063/1.3073889. URL https://doi.org/10.1063/1.3073889.

[20] Jordan W. Bye and Robin A. Curtis. Controlling phase separation of lysozyme with polyvalent anions. *The Journal of Physical Chemistry B*, 123(3):593–605, December 2018. doi: 10.1021/acs.jpcb.8b10868. URL https://doi.org/10.1021/acs.jpcb.8b10868.

[21] David Chandler. Hydrophobicity: Two faces of water. *Nature*, 417(6888):491–491, May 2002. doi: 10.1038/417491a. URL https://doi.org/10.1038/417491a.

[22] David Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647, September 2005. doi: 10.1038/nature04162. URL https://doi.org/10.1038/nature04162.

[23] Kim-Hung Chow and David M. Ferguson. Isothermal-isobaric molecular dynamics simulations with monte carlo volume sampling. *Computer Physics Communications*, 91(1-3):283–289, September 1995. doi: 10.1016/0010-4655(95)00059-o. URL https://doi.org/10.1016/0010-4655(95)00059-o.

[24] R. Clausius. Ueber eine veränderte form des zweiten hauptsatzes der mechanischen wärmetheorie. *Annalen der Physik und Chemie*, 169(12):481–506, 1854. doi: 10.1002/andp.18541691202. URL https://doi.org/10.1002/andp.18541691202.

[25] Nancy Cordoba, Laura Pataquiva, Coralia Osorio, Fabian Leonardo Moreno Moreno, and Ruth Yolanda Ruiz. Effect of grinding, extraction time and type of coffee on the physicochemical and flavour characteristics of cold brew coffee. *Scientific Reports*, 9(1), June 2019. doi: 10.1038/s41598-019-44886-w. URL https://doi.org/10.1038/s41598-019-44886-w.

[26] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117 (19):5179–5197, May 1995. doi: 10.1021/ja00124a002. URL https://doi.org/10.1021/ja00124a002.

[27] Vinícius Wilian D. Cruzeiro, Marcos S. Amaral, and Adrian E. Roitberg. Redox potential replica exchange molecular dynamics at constant pH in AMBER: Implementation and validation. *The Journal of Chemical Physics*, 149(7):072338, August 2018. doi: 10.1063/1.5027379. URL https://doi.org/10.1063/1.5027379.

[28] Christoph Dellago, Peter G. Bolhuis, and David Chandler. Efficient transition path sampling: Application to lennard-jones cluster rearrangements. *The Journal of Chemical Physics*, 108(22):9236–9245, June 1998. doi: 10.1063/1.476378. URL https://doi.org/10.1063/1.476378.

[29] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, November 2012. doi: 10.1126/science.1219021. URL https://doi.org/10.1126/science.1219021.

[30] Ken A. Dill, Sarina Bromberg, Kaizhi Yue, Hue Sun Chan, Klaus M. Ftebig, David P. Yee, and Paul D. Thomas. Principles of protein folding - a perspective from simple exact models. *Protein Science*, 4(4):561–602, December 1995. doi: 10.1002/pro.5560040401. URL https://doi.org/10.1002/pro.5560040401.

[31] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216–222, September 1987. doi: 10.1016/0370-2693(87)91197-x. URL https://doi.org/10.1016/0370-2693(87)91197-x.

[32] Peter Eastman, John Chodera, and Josh Fass. Discretizations of langevin integrator · issue nr. 2532 · openmm/openmm, 2020. URL https://github.com/openmm/openmm/issues/2532.

[33] A. Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*,

322(8):549–560, 1905. doi: 10.1002/andp.19053220806. URL https://doi.org/10.1002/andp.19053220806.

[34] D. Fennell Evans and Håkan Wennerström. *The colloidal domain: where physics, chemistry, biology, and technology meet*. Advances in interfacial engineering series. Wiley-VCH, New York, 2nd ed edition, 1999. ISBN 9780471242475.

[35] D. J. Evans. Computer "experiment" for nonlinear thermodynamics of couette flow. *The Journal of Chemical Physics*, 78(6):3297–3302, March 1983. doi: 10.1063/1.445195. URL https://doi.org/10.1063/1.445195.

[36] Denis J. Evans, William G. Hoover, Bruce H. Failor, Bill Moran, and Anthony J. C. Ladd. Nonequilibrium molecular dynamics via gauss's principle of least constraint. *Physical Review A*, 28(2):1016–1021, August 1983. doi: 10.1103/physreva.28.1016. URL https://doi.org/10.1103/physreva.28.1016.

[37] Oded Farago. Langevin thermostat for robust configurational and kinetic sampling. *Physica A: Statistical Mechanics and its Applications*, 534:122210, November 2019. doi: 10.1016/j.physa.2019.122210. URL https://doi.org/10.1016/j.physa.2019.122210.

[38] Josh Fass, David Sivak, Gavin Crooks, Kyle Beauchamp, Benedict Leimkuhler, and John Chodera. Quantifying configuration-sampling error in langevin simulations of complex molecular systems. *Entropy*, 20(5):318, April 2018. doi: 10.3390/e20050318. URL https://doi.org/10.3390/e20050318.

[39] Roberto Fernandez-Alvarez, Vladimír Ďorďovič, Mariusz Uchman, and Pavel Matějíček. Amphiphiles without head-and-tail design: Nanostructures based on the self-assembly of anionic boron cluster compounds. *Langmuir*, 34(12):3541–3554, November 2017. doi: 10.1021/acs.langmuir.7b03306. URL https://doi.org/10.1021/acs.langmuir.7b03306.

[40] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, San Diego, second edition, 2002.

[41] Andrey I. Frolov. Theory of solutions in energy representation in npt-ensemble: Derivation details, 2015.

[42] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics*. Cambridge University Press, 1902. doi: 10.1017/cbo9780511686948. URL https://doi.org/10.1017/cbo9780511686948.

[43] Raymond F. Greene and C. Nick Pace. Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, $\alpha$-chymotrypsin, and b-lactoglobulin. *Journal of*

*Biological Chemistry*, 249(17):5388–5393, September 1974. doi: 10.1016/s0021-9258(20)79739-5. URL https://doi.org/10.1016/s0021-9258(20)79739-5.

[44] Walter Grimus. 100th anniversary of the sackur-tetrode equation. *Annalen der Physik*, 525(3):A32–A35, March 2013. doi: 10.1002/andp.201300720. URL https://doi.org/10.1002/andp.201300720.

[45] Jean-Pierre Hansen and Ian R. McDonald. *Theory of simple liquids: with applications of soft matter*. Elsevier/AP, fourth edition edition, 2013. ISBN 9780123870322.

[46] Stefan Hervø-Hansen, Casper Højgaard, Kristoffer Enøe Johansson, Yong Wang, Khadija Wahni, David Young, Joris Messens, Kaare Teilum, Kresten Lindorff-Larsen, and Jakob Rahr Winther. Charge interactions in a highly charge-depleted protein. *Journal of the American Chemical Society*, February 2021. doi: 10.1021/jacs.0c10789. URL https://doi.org/10.1021/jacs.0c10789.

[47] Casper Højgaard, Christian Kofoed, Roall Espersen, Kristoffer Enøe Johansson, Mara Villa, Martin Willemoës, Kresten Lindorff-Larsen, Kaare Teilum, and Jakob R. Winther. A soluble, folded protein without charged amino acid residues. *Biochemistry*, 55(28):3949–3956, July 2016. doi: 10.1021/acs.biochem.6b00269. URL https://doi.org/10.1021/acs.biochem.6b00269.

[48] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, March 1985. doi: 10.1103/physreva.31.1695. URL https://doi.org/10.1103/physreva.31.1695.

[49] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., May 2001. doi: 10.1002/0471221317. URL https://doi.org/10.1002/0471221317.

[50] Jacob Israelachvili. *Intermolecular and surface forces*. Academic Press, Burlington, MA, 2011. ISBN 9780123919274.

[51] Satoru G. Itoh, Ana Damjanović, and Bernard R. Brooks. pH replica-exchange method based on discrete protonation states. *Proteins: Structure, Function, and Bioinformatics*, 79(12):3420–3436, October 2011. doi: 10.1002/prot.23176. URL https://doi.org/10.1002/prot.23176.

[52] Frank Jensen. *Introduction to computational chemistry*. John Wiley & Sons, Chichester, England Hoboken, NJ, 2007. ISBN 0-470-01187-4.

[53] Owen G. Jepps, Gary Ayton, and Denis J. Evans. Microscopic expressions for the thermodynamic temperature. *Physical Review E*, 62(4):4757–4763, October 2000. doi: 10.1103/physreve.62.4757. URL https://doi.org/10.1103/physreve.62.4757.

[54] Nicolas O. Johnson, Taylor P. Light, Gina MacDonald, and Yanjie Zhang. Anion–caffeine interactions studied by 13c and 1h NMR and ATR–FTIR spectroscopy. *The Journal of Physical Chemistry B*, 121(7):1649–1659, February 2017. doi: 10.1021/acs.jpcb.6b12150. URL https://doi.org/10.1021/acs.jpcb.6b12150.

[55] Elena Jordan, Felix Roosen-Runge, Sara Leibfarth, Fajun Zhang, Michael Sztucki, Andreas Hildebrandt, Oliver Kohlbacher, and Frank Schreiber. Competing salt effects on phase behavior of protein solutions: Tailoring of protein interaction by the binding of multivalent ions and charge screening. *The Journal of Physical Chemistry B*, 118(38):11365–11374, September 2014. doi: 10.1021/jp5058622. URL https://doi.org/10.1021/jp5058622.

[56] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, May 1935. doi: 10.1063/1.1749657. URL https://doi.org/10.1063/1.1749657.

[57] Christos M. Kougentakis, Lauren Skerritt, Ananya Majumdar, Jamie L. Schlessman, and Bertrand García-Moreno E. The properties of buried ion pairs are governed by the propensity of proteins to reorganize. *bioRxiv*, February 2020. doi: 10.1101/2020.02.03.932012. URL https://doi.org/10.1101/2020.02.03.932012.

[58] Georg Krainer, Timothy J. Welsh, Jerelle A. Joseph, Jorge R. Espinosa, Sina Wittmann, Ella de Csilléry, Akshay Sridhar, Zenon Toprakcioglu, Giedre Gudiškytė, Magdalena A. Czekalska, William E. Arter, Jordina Guillén-Boixet, Titus M. Franzmann, Seema Qamar, Peter St George-Hyslop, Anthony A. Hyman, Rosana Collepardo-Guevara, Simon Alberti, and Tuomas P. J. Knowles. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nature Communications*, 12(1), February 2021. doi: 10.1038/s41467-021-21181-9. URL https://doi.org/10.1038/s41467-021-21181-9.

[59] Ryan M. Kramer, Varad R. Shende, Nicole Motl, C. Nick Pace, and J. Martin Scholtz. Toward a molecular understanding of protein solubility: Increased negative surface charge correlates with increased solubility. *Biophysical Journal*, 102(8):1907–1915, April 2012. doi: 10.1016/j.bpj.2012.01.060. URL https://doi.org/10.1016/j.bpj.2012.01.060.

[60] Laura J. LaBerge and John C. Tully. A rigorous procedure for combining molecular dynamics and monte carlo simulation algorithms. *Chemical Physics*, 260(1-2):183–191, October 2000. doi: 10.1016/s0301-0104(00)00246-9. URL https://doi.org/10.1016/s0301-0104(00)00246-9.

[61] Lev Davidovič Landau and Evgeny Mikhailovich Lifshitz. *Statistical physics. 1: by E. M. Lifshitz and L. P. Pitaevskii*. Number 5 in Course of theoretical physics / L. D.

Landau and E. M. Lifshitz. Elsevier Butterworth Heinemann, Amsterdam Heidelberg, 3. ed., repr edition, 2011. ISBN 9780750633727.

[62] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics*. Springer International Publishing, 2015. doi: 10.1007/978-3-319-16375-8. URL https://doi.org/10.1007/978-3-319-16375-8.

[63] Benedict Leimkuhler and Charles Matthews. Robust and efficient configurational molecular sampling via langevin dynamics. *The Journal of Chemical Physics*, 138(17): 174102, May 2013. doi: 10.1063/1.4802990. URL https://doi.org/10.1063/1.4802990.

[64] Don S. Lemons and Anthony Gythiel. Paul langevin's 1908 paper "on the theory of brownian motion" ["sur la théorie du mouvement brownien, " c. r. acad. sci. (paris) 146, 530–533 (1908)]. *American Journal of Physics*, 65(11):1079–1081, November 1997. doi: 10.1119/1.18725. URL https://doi.org/10.1119/1.18725.

[65] Samuel Lenton, Stefan Hervø-Hansen, Anton M. Popov, Mark D. Tully, Mikael Lund, and Marie Skepö. Impact of arginine–phosphate interactions on the reentrant condensation of disordered proteins. *Biomacromolecules*, 22(4):1532–1544, March 2021. doi: 10.1021/acs.biomac.0c01765. URL https://doi.org/10.1021/acs.biomac.0c01765.

[66] Lin Li, Chuan Li, Zhe Zhang, and Emil Alexov. On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *Journal of Chemical Theory and Computation*, 9(4):2126–2136, March 2013. doi: 10.1021/ct400065j. URL https://doi.org/10.1021/ct400065j.

[67] Terence H. Lilley, Helen Linsdell, and Alfredo Maestre. Association of caffeine in water and in aqueous solutions of sucrose. *Journal of the Chemical Society, Faraday Transactions*, 88(19):2865, 1992. doi: 10.1039/ft9928802865. URL https://doi.org/10.1039/ft9928802865.

[68] Nandou Lu, Jayant K. Singh, and David A. Kofke. Appropriate methods to combine forward and reverse free-energy perturbation averages. *The Journal of Chemical Physics*, 118(7):2977–2984, February 2003. doi: 10.1063/1.1537241. URL https://doi.org/10.1063/1.1537241.

[69] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The Journal of*

*Physical Chemistry B*, 102(18):3586–3616, April 1998. doi: 10.1021/jp973084f. URL https://doi.org/10.1021/jp973084f.

[70] David C. Malaspina, Clara Viñas, Francesc Teixidor, and Jordi Faraudo. Atomistic simulations of COSAN: Amphiphiles without a head-and-tail design display "head and tail" surfactant behavior. *Angewandte Chemie International Edition*, 59(8):3088–3092, February 2020. doi: 10.1002/anie.201913257. URL https://doi.org/10.1002/anie.201913257.

[71] Vasilios I. Manousiouthakis and Michael W. Deem. Strict detailed balance is unnecessary in monte carlo simulation. *The Journal of Chemical Physics*, 110(6):2753–2756, February 1999. doi: 10.1063/1.477973. URL https://doi.org/10.1063/1.477973.

[72] F. Martin and H. Zipse. Charge distribution in the water molecule?a comparison of methods. *Journal of Computational Chemistry*, 26(1):97–105, 2004. doi: 10.1002/jcc.20157. URL https://doi.org/10.1002/jcc.20157.

[73] Glenn J. Martyna, Michael L. Klein, and Mark Tuckerman. Nosé–hoover chains: The canonical ensemble via continuous dynamics. *The Journal of Chemical Physics*, 97(4):2635–2643, August 1992. doi: 10.1063/1.463940. URL https://doi.org/10.1063/1.463940.

[74] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189, September 1994. doi: 10.1063/1.467468. URL https://doi.org/10.1063/1.467468.

[75] Glenn J. Martyna, Mark E. Tuckerman, Douglas J. Tobias, and Michael L. Klein. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87(5):1117–1157, April 1996. doi: 10.1080/00268979600100761. URL https://doi.org/10.1080/00268979600100761.

[76] Nobuyuki Matubayasi and Masaru Nakahara. Theory of solutions in the energetic representation. i. formulation. *The Journal of Chemical Physics*, 113(15):6070–6081, October 2000. doi: 10.1063/1.1309013. URL https://doi.org/10.1063/1.1309013.

[77] Nobuyuki Matubayasi and Masaru Nakahara. Theory of solutions in the energy representation. II. functional for the chemical potential. *The Journal of Chemical Physics*, 117(8):3605–3616, August 2002. doi: 10.1063/1.1495850. URL https://doi.org/10.1063/1.1495850.

[78] Nobuyuki Matubayasi and Masaru Nakahara. Theory of solutions in the energy representation. III. treatment of the molecular flexibility. *The Journal of Chemical*

*Physics*, 119(18):9686–9702, November 2003. doi: 10.1063/1.1613938. URL https://doi.org/10.1063/1.1613938.

[79] Simone Melchionna. Design of quasisymplectic propagators for langevin dynamics. *The Journal of Chemical Physics*, 127(4):044108, July 2007. doi: 10.1063/1.2753496. URL https://doi.org/10.1063/1.2753496.

[80] Frédéric Mestdagh, Arne Glabasnia, and Peter Giuliano. The brew—extracting for excellence. In *The Craft and Science of Coffee*, pages 355–380. Elsevier, 2017. doi: 10.1016/b978-0-12-803520-7.00015-3. URL https://doi.org/10.1016/b978-0-12-803520-7.00015-3.

[81] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

[82] John Mongan, David A. Case, and J. Andrew McCammon. Constant pH molecular dynamics in generalized born implicit solvent. *Journal of Computational Chemistry*, 25(16):2038–2048, December 2004. doi: 10.1002/jcc.20139. URL https://doi.org/10.1002/jcc.20139.

[83] Calvin C. Moore. Ergodic theorem, ergodic theory, and statistical mechanics. *Proceedings of the National Academy of Sciences*, 112(7):1907–1911, February 2015. doi: 10.1073/pnas.1421798112. URL https://doi.org/10.1073/pnas.1421798112.

[84] Tetsuya Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *The Journal of Chemical Physics*, 113(8):2976–2982, August 2000. doi: 10.1063/1.1287333. URL https://doi.org/10.1063/1.1287333.

[85] R. S. Mulliken. Electronic population analysis on LCAO–MO molecular wave functions. i. *The Journal of Chemical Physics*, 23(10):1833–1840, October 1955. doi: 10.1063/1.1740588. URL https://doi.org/10.1063/1.1740588.

[86] Jeffrey K. Myers, C. Nick Pace, and J. Martin Scholtz. Denaturantmvalues and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Science*, 4(10):2138–2148, October 1995. doi: 10.1002/pro.5560041020. URL https://doi.org/10.1002/pro.5560041020.

[87] Erik C. Neyts and Annemie Bogaerts. Combining molecular dynamics with monte carlo simulations: implementations and applications. *Theoretical Chemistry Accounts*, 132(2), December 2012. doi: 10.1007/s00214-012-1320-x. URL https://doi.org/10.1007/s00214-012-1320-x.

[88] P Nikunen, M Karttunen, and I Vattulainen. How would you integrate the equations of motion in dissipative particle dynamics simulations? *Computer Physics Communications*, 153(3):407–423, July 2003. doi: 10.1016/s0010-4655(03)00202-9. URL https://doi.org/10.1016/s0010-4655(03)00202-9.

[89] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera. Nonequilibrium candidate monte carlo is an efficient tool for equilibrium simulation. *Proceedings of the National Academy of Sciences*, 108(45):E1009–E1018, October 2011. doi: 10.1073/pnas.1106094108. URL https://doi.org/10.1073/pnas.1106094108.

[90] Frank Noe. Time-lagged independent component analysis (tica), 2021. URL http://docs.markovmodel.org/lecture_tica.html.

[91] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, July 1984. doi: 10.1063/1.447334. URL https://doi.org/10.1063/1.447334.

[92] C. Nick Pace, Saul Treviño, Erode Prabhakaran, and J. Martin Scholtz. Protein structure, stability and solubility in water and other solvents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1448):1225–1235, August 2004. doi: 10.1098/rstb.2004.1500. URL https://doi.org/10.1098/rstb.2004.1500.

[93] C. Nick Pace, Gerald R. Grimsley, and J. Martin Scholtz. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *Journal of Biological Chemistry*, 284(20):13285–13289, May 2009. doi: 10.1074/jbc.r800080200. URL https://doi.org/10.1074/jbc.r800080200.

[94] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, December 1981. doi: 10.1063/1.328693. URL https://doi.org/10.1063/1.328693.

[95] Coralie Pasquier, Mario Vazdar, Jan Forsman, Pavel Jungwirth, and Mikael Lund. Anomalous protein–protein interactions in multivalent salt solution. *The Journal of Physical Chemistry B*, 121(14):3000–3006, March 2017. doi: 10.1021/acs.jpcb.7b01051. URL https://doi.org/10.1021/acs.jpcb.7b01051.

[96] Philip Pearle, Brian Collett, Kenneth Bart, David Bilderback, Dara Newman, and Scott Samuels. What brown saw and you can too. *American Journal of Physics*, 78 (12):1278–1289, December 2010. doi: 10.1119/1.3475685. URL https://doi.org/10.1119/1.3475685.

[97] J. K. Percus. Approximation methods in classical statistical mechanics. *Physical Review Letters*, 8(11):462–463, June 1962. doi: 10.1103/physrevlett.8.462. URL https://doi.org/10.1103/physrevlett.8.462.

[98] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *The Journal of Chemical Physics*, 139(1):015102, July 2013. doi: 10.1063/1.4811489. URL https://doi.org/10.1063/1.4811489.

[99] Stefano Ponte. The 'latte revolution'? regulation, markets and consumption in the global coffee chain. *World Development*, 30(7):1099–1122, July 2002. doi: 10.1016/s0305-750x(02)00032-3. URL https://doi.org/10.1016/s0305-750x(02)00032-3.

[100] Rachel L. Poole and Michael G. Tordoff. The taste of caffeine. *Journal of Caffeine Research*, 7(2):39–52, June 2017. doi: 10.1089/jcr.2016.0030. URL https://doi.org/10.1089/jcr.2016.0030.

[101] Sumant Puri and Mira Edgerton. How does it kill?: Understanding the candidacidal mechanism of salivary histatin 5. *Eukaryotic Cell*, 13(8):958–964, June 2014. doi: 10.1128/ec.00095-14. URL https://doi.org/10.1128/ec.00095-14.

[102] Alan E. Reed and Frank Weinhold. Some remarks on the c–h bond dipole moment. *The Journal of Chemical Physics*, 84(4):2428–2430, February 1986. doi: 10.1063/1.450359. URL https://doi.org/10.1063/1.450359.

[103] Alan E. Reed, Robert B. Weinstock, and Frank Weinhold. Natural population analysis. *The Journal of Chemical Physics*, 83(2):735–746, July 1985. doi: 10.1063/1.449486. URL https://doi.org/10.1063/1.449486.

[104] G. Rickayzen and D. M. Heyes. A configurational temperature for molecules with hard-core or discontinuous interactions. *The Journal of Chemical Physics*, 127(14):144512, October 2007. doi: 10.1063/1.2793069. URL https://doi.org/10.1063/1.2793069.

[105] Andrea Rizzi, John Chodera, Levi Naden, Kyle Beauchamp, Patrick Grinaway, Josh Fass, Alex Wade, Bas Rustenburg, Gregory A. Ross, Andreas Krämer, Hannah Bruce Macdonald, Dominicrufa, Andy Simmonett, David W.H. Swenson, Hb0402, and Ana Silveira. choderalab/openmmtools: 0.20.0 - periodic nonequilibrium integrator, 2020. URL https://zenodo.org/record/596622.

[106] A. C. Robinson, C. A. Castaneda, J. L. Schlessman, and B. Garcia-Moreno E. Structural and thermodynamic consequences of burial of an artificial ion pair in the hydrophobic interior of a protein. *Proceedings of the National Academy of Sciences*, 111(32):11685–11690, July 2014. doi: 10.1073/pnas.1402900111. URL https://doi.org/10.1073/pnas.1402900111.

[107] J. Roche, J. A. Caro, D. R. Norberto, P. Barthe, C. Roumestand, J. L. Schlessman, A. E. Garcia, B. Garcia-Moreno E., and C. A. Royer. Cavities determine the pressure unfolding of proteins. *Proceedings of the National Academy of Sciences*, 109(18):6945–6950, April 2012. doi: 10.1073/pnas.1200915109. URL https://doi.org/10.1073/pnas.1200915109.

[108] Bradley A. Rogers, Tye S. Thompson, and Yanjie Zhang. Hofmeister anion effects on thermodynamics of caffeine partitioning between aqueous and cyclohexane phases. *The Journal of Physical Chemistry B*, 120(49):12596–12603, December 2016. doi: 10.1021/acs.jpcb.6b07760. URL https://doi.org/10.1021/acs.jpcb.6b07760.

[109] Hans Henrik Rugh. Dynamical approach to temperature. *Physical Review Letters*, 78 (5):772–774, February 1997. doi: 10.1103/physrevlett.78.772. URL https://doi.org/10.1103/physrevlett.78.772.

[110] O. Sackur. Die bedeutung des elementaren wirkungsquantums für die gastheorie und die berechnung der chemischen konstanten. *Festschrift W. Nernst zu seinem 25jährigen Doktorjubiläum gewidmet von seinen Schülern*, pages 405–423, 1913.

[111] O. Sackur. Die universelle bedeutung des sog. elementaren wirkungsquantums. *Annalen der Physik*, 345(1):67–86, 1913. doi: 10.1002/andp.19133450103. URL https://doi.org/10.1002/andp.19133450103.

[112] Shun Sakuraba and Nobuyuki Matubayasi. Ermod: Fast and versatile computation software for solvation free energy with approximate theory of solutions. *Journal of Computational Chemistry*, 35(21):1592–1608, June 2014. doi: 10.1002/jcc.23651. URL https://doi.org/10.1002/jcc.23651.

[113] Rangana Sanjeewa and Samantha Weerasinghe. Development of a molecular mechanics force field for caffeine to investigate the interactions of caffeine in different solvent media. *Journal of Molecular Structure: THEOCHEM*, 944(1-3):116–123, March 2010. doi: 10.1016/j.theochem.2009.12.027. URL https://doi.org/10.1016/j.theochem.2009.12.027.

[114] Rangana Sanjeewa and Samantha Weerasinghe. Study of aggregate formation of caffeine in water by molecular dynamics simulation. *Computational and Theoretical Chemistry*, 966(1-3):140–148, June 2011. doi: 10.1016/j.comptc.2011.02.027. URL https://doi.org/10.1016/j.comptc.2011.02.027.

[115] Marcelo M. Santoro and D. W. Bolen. Unfolding free energy changes determined by the linear extrapolation method. 1. unfolding of phenylmethanesulfonyl $\alpha$-chymotrypsin using different denaturants. *Biochemistry*, 27(21):8063–8068, October 1988. doi: 10.1021/bi00421a014. URL https://doi.org/10.1021/bi00421a014.

[116] Hugo A. F. Santos, Diogo Vila-Viçosa, Vitor H. Teixeira, António M. Baptista, and Miguel Machuqueiro. Constant-pH MD simulations of DMPA/DMPC lipid bilayers. *Journal of Chemical Theory and Computation*, 11(12):5973–5979, November 2015. doi: 10.1021/acs.jctc.5b00956. URL https://doi.org/10.1021/acs.jctc.5b00956.

[117] Christian R. Schwantes and Vijay S. Pande. Improvements in markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of Chemical Theory and Computation*, 9(4):2000–2009, March 2013. doi: 10.1021/ct300878a. URL https://doi.org/10.1021/ct300878a.

[118] Anvar Shalmashi and Fereshteh Golmohammad. Solubility of caffeine in water, ethyl acetate, ethanol, carbon tetrachloride, methanol, chloroform, dichloromethane, and acetone between 298 and 323 k. *Latin American applied research*, 40(3), July 2010.

[119] Naoya Shibayama. Allosteric transitions in hemoglobin revisited. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1864(2):129335, February 2020. doi: 10.1016/j.bbagen.2019.03.021. URL https://doi.org/10.1016/j.bbagen.2019.03.021.

[120] Seishi Shimizu. Caffeine dimerization: effects of sugar, salts, and water structure. *Food & Function*, 6(10):3228–3235, 2015. doi: 10.1039/c5fo00610d. URL https://doi.org/10.1039/c5fo00610d.

[121] Michael R. Shirts and Vijay S. Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *The Journal of Chemical Physics*, 122(14):144107, April 2005. doi: 10.1063/1.1873592. URL https://doi.org/10.1063/1.1873592.

[122] Ilan Shumilin, Christoph Allolio, and Daniel Harries. How sugars modify caffeine self-association and solubility: Resolving a mechanism of selective hydrotropy. *Journal of the American Chemical Society*, 141(45):18056–18063, October 2019. doi: 10.1021/jacs.9b07056. URL https://doi.org/10.1021/jacs.9b07056.

[123] Thomas Simonson. Free energy of particle insertion. *Molecular Physics*, 80(2):441–447, October 1993. doi: 10.1080/00268979300102371. URL https://doi.org/10.1080/00268979300102371.

[124] Thomas Soddemann, Burkhard Dünweg, and Kurt Kremer. Dissipative particle dynamics: A useful thermostat for equilibrium and nonequilibrium molecular dynamics simulations. *Physical Review E*, 68(4), October 2003. doi: 10.1103/physreve.68.046702. URL https://doi.org/10.1103/physreve.68.046702.

[125] Jayashree Srinivasan, Megan W. Trevathan, Paul Beroza, and David A. Case. Application of a pairwise generalized born model to proteins and nucleic acids: inclusion of salt effects. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(6):426–434, May 1999. doi: 10.1007/s002140050460. URL https://doi.org/10.1007/s002140050460.

[126] Thomas Steinbrecher, InSuk Joung, and David A. Case. Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations. *Journal of Computational Chemistry*, 32(15):3253–3263, August 2011. doi: 10.1002/jcc.21909. URL https://doi.org/10.1002/jcc.21909.

[127] Harry A. Stern. Molecular simulation with variable protonation states at constant pH. *The Journal of Chemical Physics*, 126(16):164112, April 2007. doi: 10.1063/1.2731781. URL https://doi.org/10.1063/1.2731781.

[128] Jason M. Swails and Adrian E. Roitberg. Enhancing conformation and protonation state sampling of hen egg white lysozyme using pH replica exchange molecular dynamics. *Journal of Chemical Theory and Computation*, 8(11):4393–4404, September 2012. doi: 10.1021/ct300512h. URL https://doi.org/10.1021/ct300512h.

[129] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, January 1982. doi: 10.1063/1.442716. URL https://doi.org/10.1063/1.442716.

[130] Charles Tanford. Isothermal unfolding of globular proteins in aqueous urea solutions. *Journal of the American Chemical Society*, 86(10):2050–2059, May 1964. doi: 10.1021/ja01064a028. URL https://doi.org/10.1021/ja01064a028.

[131] Charles Tanford. Protein denaturation. In *Advances in Protein Chemistry*, pages 1–95. Elsevier, 1970. doi: 10.1016/s0065-3233(08)60241-7. URL https://doi.org/10.1016/s0065-3233(08)60241-7.

[132] L. Tavagnacco, J. W. Brady, F. Bruni, S. Callear, M. A. Ricci, M. L. Saboungi, and A. Cesàro. Hydration of caffeine at high temperature by neutron scattering and simulation studies. *The Journal of Physical Chemistry B*, 119(42):13294–13301, October 2015. doi: 10.1021/acs.jpcb.5b09204. URL https://doi.org/10.1021/acs.jpcb.5b09204.

[133] Letizia Tavagnacco, Udo Schnupf, Philip E. Mason, Marie-Louise Saboungi, Attilio Cesàro, and John W. Brady. Molecular dynamics simulation studies of caffeine aggregation in aqueous solution. *The Journal of Physical Chemistry B*, 115(37):10957–10966, September 2011. doi: 10.1021/jp2021352. URL https://doi.org/10.1021/jp2021352.

[134] Letizia Tavagnacco, Yuri Gerelli, Attilio Cesàro, and John W. Brady. Stacking and branching in self-aggregation of caffeine in aqueous solution: From the supramolecular to atomic scale clustering. *The Journal of Physical Chemistry B*, 120(37):9987–9996, September 2016. doi: 10.1021/acs.jpcb.6b06980. URL https://doi.org/10.1021/acs.jpcb.6b06980.

[135] H. Tetrode. Berichtigung zu meiner arbeit: Die chemische konstante der gase und das elementare wirkungsquantum. *Annalen der Physik*, 344(11):255–256, 1912. doi: 10.1002/andp.19123441112. URL https://doi.org/10.1002/andp.19123441112.

[136] Mariusz Uchman, Alexei I. Abrikosov, Martin Lepšík, Mikael Lund, and Pavel Matějíček. Nonclassical hydrophobic effect in micellization: Molecular arrangement of non-amphiphilic structures. *Advanced Theory and Simulations*, 1(1):1700002, December 2017. doi: 10.1002/adts.201700002. URL https://doi.org/10.1002/adts.201700002.

[137] K. E. Van Holde, W. Curtis Johnson, and Pui Shing Ho. *Principles of physical biochemistry*. Pearson/Prentice Hall, Upper Saddle River, N.J, 2nd ed edition, 2006. ISBN 978-0-13-046427-9. OCLC: ocm57434229.

[138] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159(1):98–103, July 1967. doi: 10.1103/physrev.159.98. URL https://doi.org/10.1103/physrev.159.98.

[139] M. von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der Physik*, 326(14):756–780, 1906. doi: 10.1002/andp.19063261405. URL https://doi.org/10.1002/andp.19063261405.

[140] Xiuju Wang, Joshua William, Yucheng Fu, and Loong-Tak Lim. Effects of capsule parameters on coffee extraction in single-serve brewer. *Food Research International*, 89:797–805, November 2016. doi: 10.1016/j.foodres.2016.09.031. URL https://doi.org/10.1016/j.foodres.2016.09.031.

[141] James D. Watson and Francis H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953. doi: 10.1038/171737a0. URL https://doi.org/10.1038/171737a0.

[142] Kenneth B. Wiberg and Paul R. Rablen. Atomic charges. *The Journal of Organic Chemistry*, 83(24):15463–15469, November 2018. doi: 10.1021/acs.joc.8b02740. URL https://doi.org/10.1021/acs.joc.8b02740.

[143] Kenneth B. Wiberg and John J. Wendoloski. The electrical nature of c?h bonds and its relationship to infrared intensities. *Journal of Computational Chemistry*, 2(1):53–

57, 1981. doi: 10.1002/jcc.540020110. URL https://doi.org/10.1002/jcc.540020110.

[144] Kenneth B. Wiberg and John J. Wendoloski. Charge redistribution in the molecular vibrations of acetylene, ethylene, ethane, methane, silane and the ammonium ion. signs of the m-h bond moments. *The Journal of Physical Chemistry*, 88(3):586–593, February 1984. doi: 10.1021/j150647a051. URL https://doi.org/10.1021/j150647a051.

[145] Benjamin Widom. Some topics in the theory of fluids. *The Journal of Chemical Physics*, 39(11):2808–2812, December 1963. doi: 10.1063/1.1734110. URL https://doi.org/10.1063/1.1734110.

[146] Marcell Wolf, Felix Roosen-Runge, Fajun Zhang, Roland Roth, Maximilian W.A. Skoda, Robert M.J. Jacobs, Michael Sztucki, and Frank Schreiber. Effective interactions in protein–salt solutions approaching liquid–liquid phase separation. *Journal of Molecular Liquids*, 200:20–27, December 2014. doi: 10.1016/j.molliq.2014.08.006. URL https://doi.org/10.1016/j.molliq.2014.08.006.

[147] L.V. Woodcock. Isothermal molecular dynamics calculations for liquid salts. *Chemical Physics Letters*, 10(3):257–261, August 1971. doi: 10.1016/0009-2614(71)80281-6. URL https://doi.org/10.1016/0009-2614(71)80281-6.

[148] Amina S. Woods and Sergi Ferré. Amazing stability of the arginine-phosphate electrostatic interaction. *Journal of Proteome Research*, 4(4):1397–1402, August 2005. doi: 10.1021/pr050077s. URL https://doi.org/10.1021/pr050077s.

[149] Tahir I. Yusufaly, Yun Li, Gautam Singh, and Wilma K. Olson. Arginine-phosphate salt bridges between histones and DNA: Intermolecular actuators that control nucleosome architecture. *The Journal of Chemical Physics*, 141(16):165102, October 2014. doi: 10.1063/1.4897978. URL https://doi.org/10.1063/1.4897978.

[150] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration. *The Journal of Chemical Physics*, 100(12):9025–9031, June 1994. doi: 10.1063/1.466707. URL https://doi.org/10.1063/1.466707.

[151] Fajun Zhang, Sophie Weggler, Michael J. Ziller, Luca Ianeselli, Benjamin S. Heck, Andreas Hildebrandt, Oliver Kohlbacher, Maximilian W. A. Skoda, Robert M. J. Jacobs, and Frank Schreiber. Universality of protein reentrant condensation in solution induced by multivalent metal ions. *Proteins: Structure, Function, and Bioinformatics*, 78(16):3450–3457, September 2010. doi: 10.1002/prot.22852. URL https://doi.org/10.1002/prot.22852.

[152] Guy Ziv and Gilad Haran. Protein folding, protein collapse, and tanford's transfer model: Lessons from single-molecule FRET. *Journal of the American Chemical Society*, 131(8):2942–2947, March 2009. doi: 10.1021/ja808305u. URL `https://doi.org/10.1021/ja808305u`.

[153] Daniel Zuckerman. Everything is markovian; nothing is markovian, Jul 2015. URL `http://statisticalbiophysicsblog.org/?p=76`.

# Scientific publications

## Author contributions

### Paper I: Charge Interactions in a Highly Charge-depleted Protein

I participated in conceiving the initial idea and project, and participated in designing the research. I performed and analyzed the simulations, participated in the analysis of NMR-spectroscopy and stability data, and was main responsible for writing the manuscript.

### Paper II: Systematic Electrostatic Perturbation of a Charge-depleted Protein: Correlation between Protein Solubility and Electrostatics

I participated in conceiving the initial idea and project, and participated in designing the research. I performed and analyzed the simulations, I analyzed the solubility data, and wrote the manuscript.

### Paper III: Counter Intuitive Electrostatics upon Metal Ion Coordination to a Receptor with Two Homotopic Binding Site

I participated in designing the research and participated in analyzing the data with main responsibility for the structural analysis.

### Paper IV: Total Description of Intrinsic Amphiphile Aggregation: Calorimetry Study and Molecular Probing

I performed and analyzed the molecular dynamics simulations and participated in writing the manuscript.

### Paper v: Anion-Cation Contrast of Caffeine Solvation in Salt Solutions

I designed the research, performed and analyzed the molecular dynamics simulations for structural properties and free energetics and was main responsible writing the paper.

### Paper vi: A Surface Area Description of Salting-in and Salting-out of Caffeine

I participated in designing the research, I performed the Monte Carlo simulations and analyzed the results, and participated in writing the paper.

### Paper vii: Impact of Arginine–Phosphate Interactions on the Reentrant Condensation of Disordered Proteins

I participated in designing the research and participated in analyzing the experimental and computational data.