

**Development of Iterative  
Methods for Coarse-Graining  
Molecular Liquids**

Cover: A sodium cation (yellow) and a chloride cation (green) solvated by water. Only water molecules in the first solvation shell are shown. The configuration is a snapshot of a molecular dynamics simulation with the ECC IMC force field (see chapter 5). The water molecules are oriented with the oxygen atom towards the cation and with one of their hydrogen atoms towards the anion.

---

# Development of Iterative Methods for Coarse-Graining Molecular Liquids

---

**Entwicklung von Iterativen Vergrößerungsmethoden für die Simulation  
Molekularer Flüssigkeiten**

Vom Fachbereich Chemie der Technischen Universität Darmstadt  
Zur Erlangung des Grades Doctor rerum naturalium (Dr. rer. nat.)  
Genehmigte Dissertation von Marvin Philipp Bernhardt aus Dieburg  
Tag der Einreichung: 13. Juli 2022, Tag der Prüfung: 15. September 2022

Erstgutachter: Prof. Dr. Nico F. A. van der Vegt  
Zweitgutachter: Prof. Dr. Florian Müller-Plathe  
Darmstadt, Technische Universität Darmstadt

---



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Eduard-Zintl-Institut für  
Anorganische und  
Physikalische Chemie

---

Development of Iterative Methods for Coarse-Graining Molecular Liquids  
Entwicklung von Iterativen Vergrößerungsmethoden für die Simulation Molekularer  
Flüssigkeiten

Genehmigte Dissertation von Marvin Philipp Bernhardt

Tag der Einreichung: 13. Juli 2022

Tag der Prüfung: 15. September 2022

Darmstadt, Technische Universität Darmstadt

Bitte zitieren Sie dieses Dokument als:

URN: urn:nbn:de:tuda-tuprints-213756

URL: <http://tuprints.ulb.tu-darmstadt.de/21375>

Veröffentlicht im Jahr 2022 auf TUprints

E-Publishing-Service der TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

[tuprints@ulb.tu-darmstadt.de](mailto:tuprints@ulb.tu-darmstadt.de)

Die Veröffentlichung steht unter folgender Creative Commons Lizenz:

Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International

<https://creativecommons.org/licenses/by-sa/4.0/>

This work is licensed under a Creative Commons License:

Attribution–ShareAlike 4.0 International

<https://creativecommons.org/licenses/by-sa/4.0/>

---

## **Erklärungen laut Promotionsordnung**

### **§ 8 Abs. 1 lit. c PromO**

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

### **§ 8 Abs. 1 lit. d PromO**

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

### **§ 9 Abs. 1 PromO**

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

### **§ 9 Abs. 2 PromO**

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 13. Juli 2022

---

Marvin Philipp Bernhardt

---

---

## Erklärung zum Eigenanteil an den Veröffentlichungen

Die vorliegende Dissertation wurde der unter Leitung von Herrn Prof. Dr. Nico F. A. van der Vegt in der Zeit vom 15. Februar 2017 bis zum 31. Mai 2022 im Fachbereich Chemie der Technischen Universität Darmstadt durchgeführt. Teile dieser Arbeit sind bereits veröffentlicht.

Im Folgenden ist aufgelistet, mit welchem Anteil ich an der Veröffentlichung beteiligt war.

Mein Anteil an der folgenden Veröffentlichung beträgt 100 %

[1] M. P. Bernhardt, M. Hanke, N. F. A. van der Vegt, „Iterative Integral Equation Methods for Structural Coarse-Graining“, *The Journal of Chemical Physics* **2021**, 154, 084118

Mein Anteil an der folgenden Veröffentlichung beträgt 100 %

[2] M. P. Bernhardt, M. Hanke, N. F. A. van der Vegt, „Stability, Speed, and Constraints for Structural Coarse-Graining in VOTCA“, *eingereicht bei: Journal of Chemical Theory and Computation*

Mein Anteil an der folgenden Veröffentlichung beträgt 95 %

[3] M. P. Bernhardt, Y. Nagata, N. F. A. van der Vegt, „Where Lennard-Jones Potentials Fail: Iterative Optimization of Ion–Water Pair Potentials Based on Ab Initio Molecular Dynamics Data“, *The Journal of Physical Chemistry Letters* **2022**, 3712–3717

Mein Anteil an der folgenden Veröffentlichung beträgt 75 %

[4] M. P. Bernhardt, M. Dallavalle, N. F. A. V. der Vegt, „Application of the 2PT Model to Understanding Entropy Change in Molecular Coarse-Graining“, *Soft Materials* **2020**, 18, 274–289

---

Datum

---

Marvin Philipp Bernhardt

---

---

Die Veröffentlichungen [1], [3] und [4] sind in der Version des akzeptierten Manuskripts in diese Dissertation eingebunden. Die Veröffentlichung [2] ist in der Version des eingereichten Manuskripts eingebunden. Zu Veröffentlichung [3] wurden zusätzlich die *Supporting Information* eingebunden. Die Kapitel 3-6 entsprechen den Veröffentlichungen [1-4] in der selben Reihenfolge.

In dieser Dissertation nicht enthalten, aber im gleichen Zeitraum entstanden sind:

[5] S. Kloth, M. P. Bernhardt, N. F. A. van der Vegt, M. Vogel, „Coarse-Grained Model of a Nanoscale-Segregated Ionic Liquid for Simulations of Low-Temperature Structure and Dynamics“, *Journal of Physics: Condensed Matter* **2021**, 33, 204002

[6] M. P. Bernhardt, „DosCalc: A Parallelized Tool to Compute Degree of Freedom-Separated Density of States“, *OFS Preprints* **2022**, DOI 10.31219/osf.io/m46ph

---

---

## Erklärung zur Begutachtung der Veröffentlichungen

Weder Referent Prof. Dr. Nico F. A. van der Vegt noch Koreferent Prof. Dr. Florian Müller-Plathe der vorliegenden kumulativen Doktorarbeit waren an der Begutachtung nachstehender Veröffentlichungen beteiligt.

[1] M. P. Bernhardt, M. Hanke, N. F. A. van der Vegt, “Iterative Integral Equation Methods for Structural Coarse-Graining”, *The Journal of Chemical Physics* **2021**, *154*, 084118

[2] M. P. Bernhardt, M. Hanke, N. F. A. van der Vegt, “Stability, Speed, and Constraints for Structural Coarse-Graining in VOTCA”, *eingereicht bei: Journal of Chemical Theory and Computation*

[3] M. P. Bernhardt, Y. Nagata, N. F. A. van der Vegt, “Where Lennard-Jones Potentials Fail: Iterative Optimization of Ion–Water Pair Potentials Based on Ab Initio Molecular Dynamics Data”, *The Journal of Physical Chemistry Letters* **2022**, 3712–3717

[4] M. P. Bernhardt, M. Dallavalle, N. F. A. V. der Vegt, “Application of the 2PT Model to Understanding Entropy Change in Molecular Coarse-Graining”, *Soft Materials* **2020**, *18*, 274–289

---

Datum

---

Referent  
Prof. Dr. Nico F. A. van der Vegt

---

Koreferent  
Prof. Dr. Florian Müller-Plathe

## Acknowledgements

A chapter of my life comes to an end and it is time for thanking the people that have had part in it. I'm deeply grateful to Nico van der Vegt for his great mentoring. He has given me freedom in the choice of my projects, but also been invaluable in assessing progress and pointing me in the right direction. I also want to thank Florian Müller-Plathe for examining this thesis and the helpful discussions we had during the theory seminar. To Martin Hanke-Bourgeois, with whom I collaborated on two papers, I'm thankful for his patience and clarity that made the exchange between the mathematicians and chemists possible. I'm also thanking Rolf Schäfer and Frederik Lermyte for taking part in the evaluation committee.

Being part of a research group is a delightful experience if one has colleagues as kind as I had. Thanks for the splendid time goes to Viktor, Swami, Moritz, Madhu, Angelina, Abhishek, Vishal, Marco, David, Gregor, Divya, Ellen B., Tobias, Cahit, Francisco, Imke, Ellen P., and all the others. From the group of Florian-Müller-Plathe, I want to say thank you to Tobias, Yash, Gustavo, Andreas, Dana, and Hans-Jürgen for the collaboration on the administration of our computer clusters. A special thank you goes to David and Swami for proofreading parts of this thesis. Also mentioned shall be the people from VOTCA, especially Christoph, for helping me with my implementation. Part of my PhD has been made possible by the TRR 146, funded by the Deutsche Forschungsgesellschaft. I'm grateful for the funding and also for the opportunities to exchange with PhD students from the fields of mathematics, physics, and computer science.

I'm happy to have a family that is loving and supporting and helped me in pursuing my goals. Thank you for that, Klaus, Daniela, Björn, and Jana. I'm also very grateful for my mother Birgit, who would have liked to see me graduate. This list would be incomplete with all the wonderful people who I can call friends, those I knew before university, those that I met here, and all those that found their way into my life on other ways. And it would also be incomplete without Verena, who has given me support and motivation when I needed it most. Thank you!



---

---

# Abstract

---

The enormous number of atoms in biological and macromolecular systems can prohibit the direct application of atomistic molecular dynamics (MD) simulations. This limitation motivates the construction of coarser models for molecular systems. Bottom-up coarse-graining methods derive potentials between coarse-grained (CG) beads, which represent groups of atoms, by matching properties of a reference atomistic force field. By developing reliable and well-understood models, time and length scales inaccessible by atomistic MD simulations become reachable. Finding effective pair potentials that represent a certain radial distribution function (RDF) is an inverse problem that needs to be solved iteratively, e.g. by Newton's method. In every iteration, a potential is used in an MD simulation to calculate the RDF. From the mismatch of the RDF and the target RDF, a potential update is calculated and a new iteration starts. The same technique can also be used to obtain atomistic force fields from ab initio MD simulations. Besides the primary challenge of matching structure, the dynamic and thermodynamic properties are altered when changing resolution. It is of interest how those properties are changing and how some of them might be retained.

The first topic in this work is the use of integral equation theory for bottom-up coarse-graining. The theory provides an approximate link between structure and potential, e.g. via the reference interaction site model (RISM) and the hypernetted chain (HNC) closure relation. This link can be used to provide a good initial guess for the pair potential and an approximation to the Jacobian matrix for an iteration in Newton's method. While the exact Jacobian is in principle accessible from sampling certain covariances, as done in the inverse Monte Carlo (IMC) method, the HNC Newton's method is distinctly faster. The integral equation coarse-graining theory is in two steps generalized, such that is finally applicable to any molecular mixture. Instabilities in the iterative RDF matching process are examined and a modification for their avoidance is developed. By changing from a Newton to a Gauss-Newton method, constraints can be included in the potential updates. Thermodynamic constraints, such as pressure, osmotic pressure in implicit solvent models, and the enthalpy of vaporization, are developed and the combination of multiple constraints is explored. All methodological advancements are implemented in the open-source coarse-graining software package VOTCA.

Secondly, iterative methods are applied to derive ion-water pair potentials from AIMD data. Instead of using a fixed parametric form for the ion-water potential, a free-form tabulated potential is derived. By comparing the derived potentials with the

---

---

parametric Lennard-Jones (LJ) form, which is typically used in electrolyte force fields, it is found that the latter has an overly steep repulsion flank. This directly affects dynamical properties such as vibrational frequencies of the ions in the solvation shell. With the derived potentials, experimental frequencies are closely matched, while the LJ potentials fail to do so. Also, the solvation entropy is in better agreement with experimental values when departing from the LJ form.

Thirdly, the dynamic and thermodynamic effects of coarse-graining molecular liquids with different levels of resolution are assessed. By comparing the vibrational density of states (VDOS) of a mapped atomistic trajectory with that of a derived CG model, acceleration of translational and rotational dynamics as well as washing out of vibrational dynamics with decreasing resolution are made visible. The two-phase thermodynamic model is used to connect the VDOS of the liquid systems to their entropy. This allows for a detailed investigation into the contributions to the entropy and how they change with the resolution of the CG model. The loss in entropy is found to happen in several steps, where the loss of rotational degrees of freedom plays a larger role than the loss of vibrations.

---

# Zusammenfassung

---

Die enorme Anzahl an Atomen in biologischen und makromolekularen Systemen kann die direkte Anwendung atomar-aufgelöster Molekulardynamiksimulationen (MD) verhindern. Diese Einschränkung motiviert die Konstruktion vergrößerter (engl. coarse-grained) Modelle für molekulare Systeme. Systematische Vergrößerungsmethoden leiten Potenziale zwischen Kugeln, welche Gruppen von Atomen representieren, ab, um Eigenschaften eines atomistischen Referenzmodells widerspiegeln. Durch die Entwicklung zuverlässiger und gut verstandener Modelle können Zeit- und Längenskalen erreicht werden, die für atomar-aufgelöste MD-Simulationen nicht erreichbar sind. Die Bestimmung effektiver Paarpotentiale, die eine bestimmte radiale Verteilungsfunktion (engl. radial distribution function, RDF) erzeugen, ist ein inverses Problem, das iterativ gelöst werden muss, z. B. mit dem Newtonverfahren. In jeder Iteration wird ein Potenzial in einer MD-Simulation verwendet, um die RDF zu berechnen. Aus dem Unterschied zwischen RDF und Ziel-RDF wird eine Verbesserung des Potenzials berechnet und eine neue Iteration beginnt. Die gleiche Technik kann auch verwendet werden, um klassische atomistische Modelle aus ab initio MD-Simulationen (AIMD) zu erhalten. Neben der primären Herausforderung der Strukturanpassung ändern sich die dynamischen und thermodynamischen Eigenschaften, wenn sich die Auflösung verändert. Es ist von Interesse zu wissen, wie sich diese Eigenschaften verändern und wie einige von ihnen fixiert werden können.

Das erste Thema in dieser Arbeit ist die Verwendung der Integralgleichungstheorie im Kontext systematischer Vergrößerungsmethoden. Die Theorie bietet eine approximative Verbindung zwischen Struktur und Potenzial zum Beispiel über das *reference interaction site model* (RISM) und der *hypernetted-chain* (HNC) Gleichung. Diese Verbindung kann verwendet werden, um eine gute Anfangsschätzung des Paarpotentials und eine Näherung der Jacobi-Matrix für eine Iteration im Newtonverfahren zu erhalten. Während die exakte Jacobi-Matrix im Prinzip aus Trajektorien berechenbar ist, wie es in der inversen Monte-Carlo-Methode (IMC) genutzt wird, ist das HNC-Newtonverfahren deutlich schneller. Die Integralgleichungstheorie-basierten Vergrößerungsmethoden werden in zwei Schritten verallgemeinert, sodass sie schließlich auf jedes molekulare System anwendbar sind. Instabilitäten in dem neuen Newtonverfahren werden untersucht und eine Modifikation zu deren Vermeidung entwickelt. Durch den Übergang vom Newton- zu einem Gauß-Newton-Verfahren können Beschränkungen (engl. constraints) für die Potenzialupdates einbezogen werden. Thermodynamische Beschränkungen wie zum Beispiel des Druckes, des osmotischen Druckes in Modellen mit impliziten Lösungsmitteln und der

---

---

Verdampfungsenthalpie werden erforscht und die Kombination mehrerer Beschränkungen untersucht. Alle methodischen Fortschritte wurden in das Open-Source-Softwarepaket VOTCA implementiert.

Zweitens werden iterative Methoden angewandt, um Ionen-Wasser-Potenziale aus AIMD-Daten abzuleiten. Anstatt eine feste parametrische Form für das Ionen-Wasser-Potential zu nutzen wird ein flexibles tabuliertes Potenzial optimiert. Ein Vergleich der optimierten Potenziale mit der parametrischen Lennard-Jones-Form (LJ), die üblicherweise in Elektrolyt-Kraftfeldern verwendet wird, zeigt, dass das LJ-Potenzial eine zu steile Abstoßungsflanke aufweist. Dies wirkt sich direkt auf die dynamischen Eigenschaften wie die Schwingungsfrequenzen der Ionen in der Hydrathülle aus. Mit den abgeleiteten Potenzialen werden experimentelle Frequenzen gut wiedergegeben, während dies bei den LJ-Potenzialen nicht der Fall ist. Auch die Hydrationsentropie stimmt besser mit den experimentellen Werten überein, wenn von der LJ-Form abgewichen wird.

Drittens werden die Auswirkungen verschiedener starker Vergrößerungen auf die Dynamik und Thermodynamik molekularer Flüssigkeiten untersucht. Durch den Vergleich der Schwingungszustandsdichte (engl. vibrational density of states, VDOS) werden die Beschleunigung der Translations- und Rotationsdynamik sowie das Auswaschen der Schwingungsdynamik mit zunehmender Vergrößerung sichtbar gemacht. Das *two-phase thermodynamic model* wird verwendet, um die VDOS der flüssigen Systeme mit ihrer Entropie zu verknüpfen. Dies ermöglicht eine detaillierte Untersuchung der Beiträge zur Entropie und wie sie sich mit der Auflösung des CG-Modells ändern. Es wird festgestellt, dass der Entropieverlust in mehreren Schritten erfolgt, wobei der Verlust von Rotationsfreiheitsgraden eine größere Rolle spielt als der Verlust von Schwingungen.

---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Theory</b>	<b>9</b>
2.1	Newton-Raphson and Gauss-Newton Method . . . . .	9
2.2	Bottom-up Coarse-Graining . . . . .	13
2.3	Integral Equation Theory . . . . .	21
<b>3</b>	<b>Iterative integral equation methods for structural coarse-graining</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	Theory . . . . .	32
3.3	Method . . . . .	37
3.4	Results and Discussion . . . . .	41
3.5	Conclusion and Outlook . . . . .	51
<b>4</b>	<b>Stability, Speed, and Constraints for Structural Coarse-Graining in VOTCA</b>	<b>55</b>
4.1	Introduction . . . . .	56
4.2	Theory . . . . .	57
4.3	Simulation Details . . . . .	71
4.4	Results . . . . .	75
4.5	Conclusion . . . . .	87
<b>5</b>	<b>Where Lennard-Jones Potentials Fail: Iterative Optimization of Ion-Water Pair Potentials Based on Ab Initio Molecular Dynamics Data</b>	<b>89</b>
5.1	Letter . . . . .	90
5.2	Supporting Information . . . . .	97
<b>6</b>	<b>Application of the 2PT model to understanding entropy change in molecular coarse-graining</b>	<b>119</b>
6.1	Introduction . . . . .	119
6.2	Theoretical Background . . . . .	122
6.3	Simulation Details . . . . .	124
6.4	Results and Discussion . . . . .	126
6.5	Conclusion . . . . .	139
<b>7</b>	<b>Conclusions and Outlook</b>	<b>141</b>
	<b>Reproducibility</b>	<b>147</b>
	<b>Bibliography</b>	<b>149</b>

---

---

## Abbreviations

MD	molecular dynamics
FG	fine-grained
CG	coarse-grained
RDF	radial distribution function
IBI	iterative Boltzmann inversion
IMC	inverse Monte Carlo
RE	relative entropy
CRW	conditional reversible work
PMF	potential of mean force
m-PMF	many-body potential of mean force
DOF	degree of freedom
VDOS/DOS	vibrational density of states
LJ	Lennard-Jones [potential]
HNC	hypernetted-chain [equation]
IE	integral equation [theory]
OZ	Ornstein-Zernike [equation]
RISM	reference interaction site model
PY	Percus-Yevick [equation]
MSA	mean spherical approximation
DFT	density functional theory
AIMD	ab initio molecular dynamics
2PT	two-phase thermodynamic [method]
TI	thermodynamic integration

---

# 1 Introduction

---

The simulation of molecular systems with computers has become an irreplaceable tool in science and engineering.[1–4] In computer simulations of chemical systems, one can do things that are difficult to do in reality: control and inspect every degree of freedom (DOF), investigate dangerous or not-yet-synthesized substances, and freely enforce conditions that are challenging to experimentalists. Interestingly, the predictive power of computational models is often not limited by the understanding of fundamental physics. Quite the opposite, there is little expected effect on the field of computational chemistry from further advances in particle physics or quantum mechanics. Instead, the limitation in computational resources prohibits using simulations with full detail at time and length scales that are interesting for certain systems of interest. An example is the simulation of a whole biological cell, which, so far, has not been performed at atomistic resolution due to the immense number of atoms.[5] Another example is polymer chains in melts or solutions, which have length scales of tens of nanometers and take seconds for equilibration.[1] The limitation in computational resources necessitates simplified models for the whole system or parts of it. These models hide details that are assumed to be less critical to the properties of interest. Therefore, the systematic design and parametrization of the simplified models are crucial objectives in the field of molecular simulations.

The creation of **classical atomistic force fields** for molecular dynamics (MD) simulations is a typical example of the parametrization problem of a simplified model. Four force fields that are popular today, CHARMM, OPLS, AMBER, and GROMOS, have been developed in the 1980s and 1990s.[6–9] The creators of those force fields had to find ways to model interactions between atoms that are, in reality, governed by quantum mechanics. Due to the poor performance of classical computers for quantum calculations, a fact first noted by Richard Feynman[10], simplified potentials that hide all electronic DOFs were used. Today, most atomistic force fields, including the four mentioned above, use Lennard-Jones (LJ) potentials and a Coulomb term for the intermolecular interactions with the atoms' nuclei as the interaction centers. This combination of potentials is very successful, but it neglects phenomena like polarization or chemical reactions. Extensions to atomistic force fields have been developed which can overcome those deficiencies to some degree.[11, 12] For the parametrization of atomistic force fields' intramolecular potentials, quantum mechanical calculations were used to determine bond, angle, and dihedral parameters. The intermolecular potential parameters were mostly fitted to match experimental values of liquid density and enthalpy of vaporization.[7–9]

---

---

Particularly interesting is the situation of **aqueous electrolyte force fields**, which contain strong electrostatic interactions compared to most organic solvents. For pure water alone, there exist a large number of models consisting of three, four, or five interaction sites.[13–15] An important aspect is the effect of polarization in water. Two kinds of polarization exist in liquid water, one from the reorientation of the water molecules and one from the distortion of the molecule’s electrons. In non-polarizable models, the missing electronic polarization is accounted for by having a higher dipole moment than predicted from experimental gas-phase measurements.[16] Till approx 14 years ago, the charge of ions was usually taken as multiples of 1 e. It was the idea of Leontyev and Stuchebrukhov to scale the charge of ions to account for the missing electronic polarization of the water molecules that would reduce the cation-anion attraction.[17] Scaled (in water usually by a factor of 0.75) charges have been found to solve issues of low solubility and erroneous adsorption of ions to proteins.[18–20] Apart from charge scaling, a lot of work has gone into the parametrization of LJ parameters on properties like solvation free energy, solvation entropy, and coordination number.[21, 22] For salts with differently sized cation and anion, it is necessary to diverge from common LJ combination rules usually used to obtain parameters for the cation-anion interaction.[23] Few studies have derived ion-water interactions from quantum mechanical calculations instead of using experimental values. Tóth has calculated parametrized ion-oxygen and ion-hydrogen pair potentials from Møller-Plesset perturbation theory calculations and obtained realistic radial distribution functions (RDFs) when using the pair potentials in MD simulations.[24] Lyubartsev et al. found that short-ranged pair potentials derived from ab initio MD (AIMD) are best described by an exponential function, not by an LJ potential.[25] However, it is unclear in what regard electrolytes are better represented with ion-water interactions derived from quantum calculations compared to commonly used force fields. The systematic derivation of potentials bottom-up from detailed electronic calculations opens a second route to electrolyte force fields that is independent of the LJ potential form. So far, studies suggest that the short-range repulsion is different, but what effect that has on electrolyte properties is unexplored.

Despite an ongoing increase in computational performance, being it from more processing power or better software[26], some systems are orders of magnitude too large for atomistic MD. This limitation motivates more simplified models of molecular systems. The term **coarse-graining** describes different methods that reduce the number of DOFs relative to atomistic models. The aforementioned computational limitation gets tackled by this in three ways:

1. the reduced number of DOFs directly reduces the computational load,
2. less friction between molecules leads to faster diffusion and thereby faster equilibration, and
3. the removal of fast DOFs allows for a larger integration time step.[27]

An early appearance of this idea is that of a united-atom force field in which the nonpolar

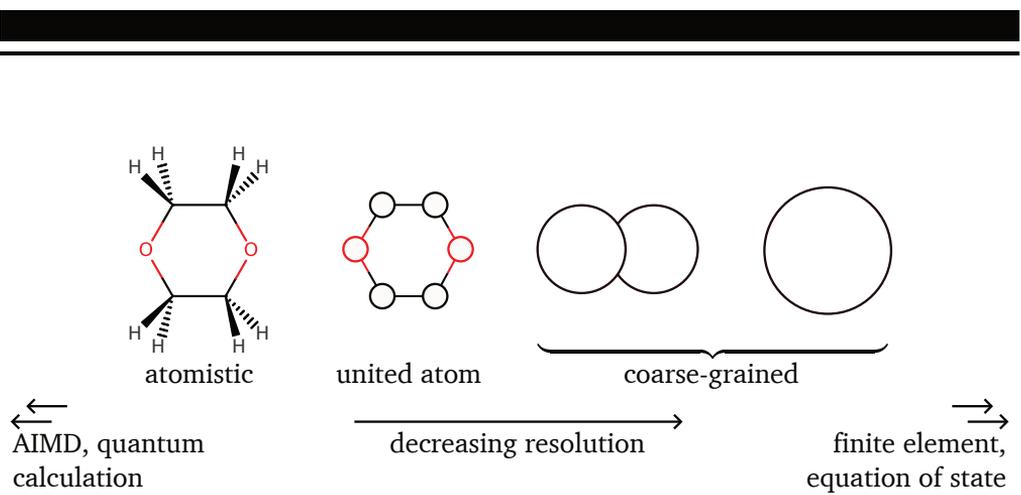


Figure 1.1: A comparison of 1,4-dioxane models at four different spatial resolutions. Atomistic and united-atom models have a well-defined resolution while for CG models it can vary. At higher resolutions, electrons are part of the model, and quantum calculations are used. At much lower resolutions, condensed matter can, e.g., be simulated with finite element methods or described by an equation of state.

hydrogen atoms are fused with their parent carbon atom.[28] Since hydrogen atoms make up a large proportion (by number) of the atoms in organic matter and their fast vibrations require small time steps, removing them from the simulation reduces computational load.[6] The speedup through united-atom force fields is still insufficient for soft matter systems; therefore, coarser representations are needed.[29] Coarse-grained (CG) models represent molecules with beads (sometimes also called super-atoms), where every bead represents a group of atoms. The difference between atomistic, united-atom, and CG resolution is portrayed in fig. 1.1. The number of atoms per bead differs significantly between models and can be as high as hundreds of atoms per bead.[30] The mapping operator formally defines the relation between the atomistic and the CG levels of representation. Its design is, so far, mostly an intuitive choice of the researcher; also, some effort has been made to investigate the effects of different mapping schemes. Foley et al. found, using the simplified Gaussian network model, that for proteins the *information efficiency* peaks at around 1 bead per 20 amino acids.[31] Dallavalle et al. have derived pair potentials for n-dodecane at different resolutions and found lower resolutions to less accurately reproduce thermodynamic properties like surface tension and thermal expansion coefficient.[32] Chakraborty et al. found that for liquid alkanes symmetry-preserving mappings have no inherent advantage over other symmetry-breaking mappings.[33] Many aspects of the choice of the mapping operator remain unclear. It is logical that with decreasing resolution the number of DOFs and therefore also the entropy decreases. But how fast the entropy decreases and how much of that can be attributed to

---

---

the mapping alone or the CG potential is open to question.

Coarse-grained MD simulations have found **applications** in different fields in which time and length scales prevent the use of atomistic MD simulations. The expression *multi-scale modeling* is used to express that features at one scale are important to properties of interest at a different scale. Reith et al. have used CG models of poly(acrylic acid) for the prediction of hydrodynamic radii which are in accordance with light scattering experiments.[34] A CG model of multi-layer graphene has allowed an analysis of its fracture and elastic behavior, preserving the anisotropy from the hexagonal structure.[35] For ionic liquids, CG models can predict structural and dynamical features at low temperatures which are difficult to equilibrate with FG models.[36] CG Models have also been used in computational biology, e.g. to study the self-assembly of membrane proteins.[37] Further examples of applications can be found in the review papers by Ingólfson et al. as well as Joshi and Deshmukh.[38, 39]

Similar to the parametrization of atomistic force fields, the **parametrization of CG models** either goes *bottom-up* from fine-grained (FG) models or *top-down* using experimental observables. Historically, the parametrization of CG models has mostly been bottom-up, i.e., deriving parameters from reference FG force fields. Bottom-up coarse-graining has the advantage of being systematic: a reparametrization for different state points or compositions is straightforward.[40] A slight disadvantage of bottom-up coarse-graining is that the reference model's accuracy limits the CG model's accuracy. The opposite approach, deriving parameters top-down from experimental observables such as partitioning free energies between polar and nonpolar phases, has also been used. It led, for example, to the semi-quantitative Martini force field.[41, 42]

Bottom-up parametrization of CG models can target many properties of the reference system.[43] In structural coarse-graining, the CG model is constructed to represent certain structural features of the reference system. For liquids, a typical target is the bead-bead RDF. But it encompasses also intramolecular structure measures such as bond and angle distributions between beads. Another class of methods is based on matching the forces between beads. The idea was first conceived for deriving potentials between atoms from *ab initio* calculations.[44] The *multiscale coarse-graining method* by Izvekov and Voth is based on this idea.[45] In it, pair potentials are optimized to yield the same total force on a mapped bead as the associated group of atoms experiences in the unmapped FG configuration. Another method that is also formulated in terms of forces, is *effective force coarse graining*. [46] It computes the direct forces between groups of atoms and projects them on the connecting vector. A method that is not directly based on forces is the *conditional reversible work* (CRW) method. It matches the pair-wise reversible work it takes to approach two groups of atoms that represent two beads.[47]

The particular problem of parametrizing pair potentials to match the RDF is an **inverse problem**. The problem is illustrated in fig. 1.2. For single-component systems, the Henderson Theorem states that there is a one-to-one mapping between RDF and pair potential (plus an arbitrary constant).[48, 49] Consequently, the problem of finding a pair

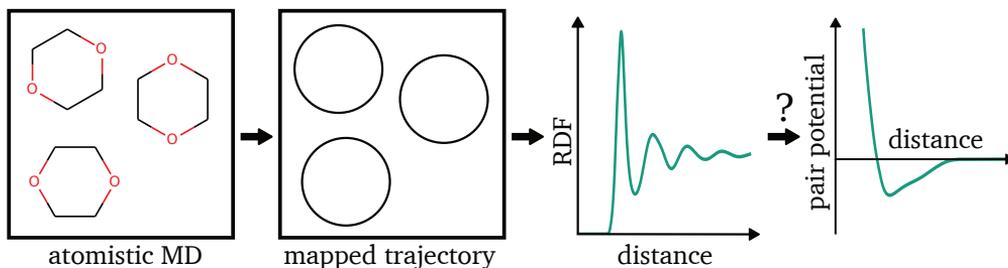


Figure 1.2: The process of structural coarse-graining and the contained inverse problem for an example system of 1,4-dioxane. An atomistic MD simulation is run as a reference, a one-bead-per-molecule mapping is applied, and the RDF is calculated from the mapped configurations. No closed-form function exists from the RDF to a CG pair potential that reproduces the RDF in an MD simulation. In the opposite direction, one can run an MD simulation with a given potential to obtain the RDF. These properties make the determination of the pair-potential an inverse problem.

potential that matches the RDF has one unique solution. However, this finding alone does not implicate that a directly computable function from the RDF to the pair potential exists. It is an inverse problem because in the opposite direction the RDF can be obtained easily from a pair potential via an MD simulation. A typical method for the solution of an inverse problem is Newton-Raphson or Newton's method. The *inverse Monte Carlo* (IMC) method is Newton's method in which the Jacobian matrix is sampled from cross-correlations of particle numbers at different distances.[50] The IMC method has some problems in stability and therefore requires a good starting point for its iterations. Another method that attempts to match the RDF and has been applied frequently is *iterative Boltzmann inversion* (IBI).[51, 52] It has, in contrast to IMC, very high stability but converges slowly.[53, 54] The conditions under which it does converge are unclear, although some progress has been made in that direction.[55] A further structural coarse-graining method is the relative entropy (RE) framework which maximizes the distributional overlap between the FG and the CG configurational probability density function.[56] It has a very flexible definition and can also be used for applications other than pair potentials, such as a mean-field model for a lattice gas.[57] While IMC and IBI both optimize the CG pair potential to match the RDF, they both have their problems. There is a need for methods that are fast, stable, and reliably converge to the unique potential that the Henderson theorem predicts. In principle, IMC represents Newton's method and should provide these properties, but its instability is not understood and limits its application.

The change in the resolution through coarse-graining is not without side effects. What is often called issues of **representability** encompasses differences in properties

---

between FG and CG models that the CG model was not optimized for. Through the loss of fast DOFs, friction between molecules is reduced and diffusion is accelerated. While this is an advantage if one is only interested in sampling configurations it can be a disadvantage. The loss in friction can be compensated by reintroducing friction in the CG simulation or by predicting the speedup and accounting for it after the simulation.[58–60, 281] Another problem, one that is most pronounced with structural coarse-graining, is the representation of thermodynamic properties.[43] The classical example is the pressure of a CG model. CG potentials computed with IBI or IMC often show very high pressures.[27] Under  $NpT$  conditions, those pair potentials would lead to evaporation, whereas the reference FG model predicts a liquid. The IBI method has been extended to p-IBI, where after each iteration a ramp potential is added to the new potential to match the pressure.[52, 61] A similar approach (KB-IBI) has been developed to match the Kirkwood-Buff integrals of CG models.[62] For force-matching, a constraint exists, that fixes the instantaneous virial and thereby ensures the right pressure.[45] Some approaches go beyond modifying the pair potentials. Dunn et al. have made compelling arguments, for the extension of the CG Hamiltonian with a volume-dependent term.[63, 64] They argue that a CG potential behaves more like a free energy than an energy because it contains entropic contributions from the FG configurations mapped into a single CG configuration. Therefore, CG potentials should depend on temperature and volume. Without those extensions to the Hamiltonian, i.e. for pure pair potentials, it is uncertain which thermodynamic properties can be represented. Furthermore, the ramp corrections ansatz, as used in p-IBI and KB-IBI, can not be adapted to match multiple properties in one potential.[62]

The view of CG potentials as free energies has also implications on the **transferability** of CG models. In this context, a perfectly transferable model would be one that would be valid at state points different from the one it was derived at. In general, one can not expect CG models to be transferable. As will be shown in the next chapter, the *ideal* CG potential, the many-body potential of mean force (m-PMF), is temperature-dependent. Approximate CG potentials are also found to be varying when derived at different temperatures.[65] A simple measure to obtain CG pair potentials at different temperatures is to scale them with the temperature, although the scaling is not consistent between different molecules.[66] The aforementioned volume-dependent CG Hamiltonians are also not per se temperature-independent but can be inter and extrapolated from parametrizations at different temperatures.[67] While it is clear, that CG pair potentials alone are not very temperature transferable, the prospect is less bleak for variations in composition. For example, Villa et al. have shown, that the transferability of a CG model of benzene water mixtures depends on the method of parametrization as well as the concentration at which the model is derived.[68]

Recently, **integral equation theory** has been used in various ways in coarse-graining methods. Integral equation (IE) theory itself is older than computer simulations and enables, much like MD simulations, a connection between microscopic interactions and

---

---

macroscopic properties. In the field of coarse-graining, it is useful because it provides approximate equations that connect the pair potential and the RDF. Guenza et al. used IE theory to derive an analytical solution for CG potentials of polymer melts at very low resolution.[69, 70] Wang et al. have used IE theory to obtain CG pair potentials for polyethylene glycol monomers which are then used in a CG model of oligomers.[71, 72] They also used their pair potentials as an initial guess for IBI and further improved their model. Mashayak et al. have extended this idea to mixtures.[73] Finally, Levesque et al. and Delbary et al. have derived methods to iteratively improve pair potentials using IE theory.[74, 75] Both have derived and applied their methods only to monoatomic liquids. In the context of the aforementioned need for alternatives or improvements to IBI and IMC, it would be interesting if IE-based methods can fill this gap. For molecular liquids and mixtures with arbitrary mapping schemes, the existing methods are not applicable and need further development.

From the above considerations the following questions can be distilled:

1. Can a coarse-graining method for RDF matching based on IE theory meet the need for better speed and stability?
2. Can such an IE-based method be generalized to arbitrary systems?
3. How can multiple thermodynamic properties be represented in pure CG pair potentials without resorting to a ramp correction?
4. Can iterative coarse-graining methods systematically derive interaction potentials between water molecules and ions from AIMD data.
5. What properties are sensitive to the short-range form of the ion-water interaction potential and possibly better represented by AIMD
6. What can be learned about changes in dynamics and thermodynamics at different CG resolutions and the contributions from the CG potential.

This introduction is followed by five chapters. In chapter 2, the basics of Newton's method, the theory of coarse-graining, and an introduction to IE theory are presented. Chapter 3 explores the application of IE theory to iterative coarse-graining and applies them to symmetric molecules and implicit solvent models. It is the continuation and expansion of the work of Delbary et al and answers question 1. In chapter 4 the IE theory-based coarse-graining methods are generalized, accelerated, and different thermodynamic constraints introduced. The aspect of numerical stability in IBI, IMC, and the IE-based methods is addressed and questions 2 and 3 are dealt with. An application of iterative methods is presented in chapter 5 where they are used to derive short-range ion-water potentials from AIMD simulation data. The derived potentials are compared to the widely used LJ potential in terms of various properties and questions 4 and 5 are answered.

---

---

Finally, in chapter 6, the effects of the mapping scheme and the CG potential on dynamics and thermodynamics are examined. To that end, an approximative model is used that connects the vibrational density of states with the thermodynamic entropy, and question 6 is addressed.

---

## 2 Theory

---

The basics of molecular dynamics (MD) simulations will not be covered in this chapter and the reader is referred to the literature.[76, 77] Instead, a concise introduction to three distinct topics central to this work will be given.

### 2.1 Newton-Raphson and Gauss-Newton Method

In this section, a set of numerical root-finding and optimization methods are introduced. Newton's method and the Gauss-Newton method with constraints are explained with short examples. Those methods are at the core of some coarse-graining methods discussed later.

The Newton-Raphson method, also called Newton's method, is an iterative method that can be used to find function roots (zeros). The method takes an initial guess for a parameter  $p$  and repeatedly improves it such that the function  $f(p)$  approaches zero. The iteration  $k$  for finding the root of a differentiable function  $f(p)$  in one dimension is written

$$p_{k+1} = p_k - \underbrace{\frac{f(p_k)}{f'(p_k)}}_{\Delta p_k}. \quad (2.1)$$

An example is calculating the numerical value of  $\sqrt{2}$ . It is an inverse problem because there is no single-step function to calculate the square root of a number, but squaring a number is straightforward (in a computer). Finding the value of  $\sqrt{2}$  is equivalent with finding the positive root of  $f(p) = p^2 - 2$ . Newton' method requires the derivative of  $f$

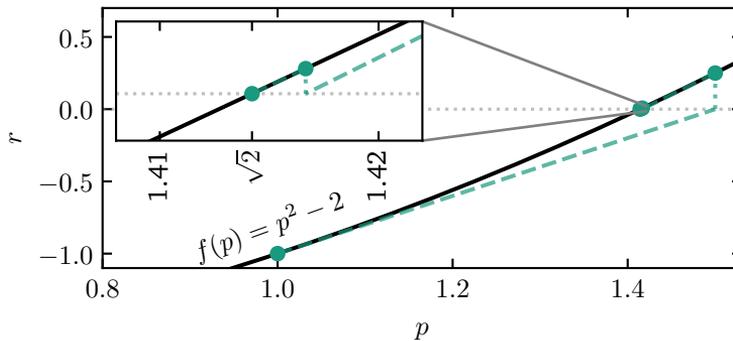


Figure 2.1: Three steps in the Newton-Raphson method for finding the value of  $\sqrt{2}$ . The dashed lines visualize how the new value  $p_{k+1}$  is obtained from the value and derivative of the function at  $p_k$ . The vertical dotted lines show the value of  $f(p_{k+1})$ .

which is given by  $f'(p) = 2p$ . Applying eq. (2.1) and starting with  $p_0 = 1$  one obtains

$$\begin{aligned}
 p_1 &= p_0 - \frac{f(1)}{f'(1)} = 1 - \frac{-1}{2} = 1.5 \\
 p_2 &= p_1 - \frac{f(1.5)}{f'(1.5)} = 1.5 - \frac{0.25}{3} = 1.41\bar{6} \\
 p_3 &= p_2 - \frac{f(1.41\bar{6})}{f'(1.41\bar{6})} = 1.41\bar{6} - \frac{0.0069\bar{4}}{2.8\bar{3}} = 1.41421568\dots \\
 &\vdots \\
 \sqrt{2} &= 1.41421356\dots
 \end{aligned}$$

After three iterations, the first six digits are correctly obtained. The process is depicted in fig. 2.1 which also shows the geometrical motivation behind Newton's method: the point  $(p_k, f(p_k))$  and the slope  $f'(p_k)$  determine a tangent line and  $p_{k+1}$  marks the spot where the line crosses the x-axis.

Newton's method converges quadratically in the vicinity of the root. In other words, in each iteration, the relative distance between  $p$  and the root is squared. However, Newton's method is not guaranteed to find a solution as it can get stuck in loops or diverge. One can also use Newton's method to find an extremum of a function, i.e., a point where the function has a derivative of zero. To find this point, one can use eq. (2.1) with  $f(p) = g'(p)$ , where  $g$  is the twice differentiable function to be minimized or maximized.

---

Newton's method generalizes to functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  with  $n$  input parameters  $\mathbf{p} = p_1, p_2, \dots, p_n$  and  $n$  output parameters  $\mathbf{r} = r_1, r_2, \dots, r_n$ . In each iteration, the new parameters are obtained by

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \underbrace{\mathbf{J}_k^{-1} \mathbf{r}_k}_{\Delta \mathbf{p}_k}. \quad (2.2)$$

The matrix  $\mathbf{J}$  is the Jacobian with elements  $J_{i,j} = \frac{\partial r_i}{\partial p_j}$ . The subscript  $k$  in  $\mathbf{J}_k$  denotes that the derivative is taken at the value of the parameters  $\mathbf{p}_k$  at iteration  $k$ , i.e.  $\mathbf{J}_k = \mathbf{J}(\mathbf{p}_k)$ .

For some problems, there is no exact root to be found, for example if the problem is overdetermined and more equations than parameters exist. Newton's method can not be applied directly, since it only works with an invertible square Jacobian matrix. Instead of finding a root, the aim for such problems is to minimize  $m$  residuals  $\mathbf{r} = r_1, r_2, \dots, r_m$  by variation of  $n$  parameters  $\mathbf{p}$  with  $m \geq n$ . A typical example is data fitting, where a function with few parameters approximates measured data points. In that case, the residual  $r_i$  is the distance of data point  $i$  from the fitting function's value at the same point. The Gauss-Newton method is similar to Newton's but solves the least squares problem of minimizing  $\sum_i^n r_i^2$ . Its iterative update is given by

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \underbrace{(\mathbf{J}_k^T \mathbf{J}_k)^{-1} \mathbf{J}_k^T \mathbf{r}_k}_{\Delta \mathbf{p}}. \quad (2.3)$$

Note that eqs. (2.1) and (2.3) are explicit formulations of the update, while numerically it is advantageous to solve linear equation systems without inverting matrices. The iterative Gauss-Newton update can be constrained to fulfill linear constraints of the form

$$\mathbf{K} \Delta \mathbf{p} = \mathbf{d} \quad (2.4)$$

where  $\mathbf{K}$  is a  $n_c \times n$  matrix and  $\mathbf{d}$  an  $n_c$  element vector with  $n_c$  being the number of constraints. For a single constraint,  $\mathbf{K}$  becomes a vector, and  $\mathbf{d}$  is a single number. Depending on the constraint, both  $\mathbf{K}$  and  $\mathbf{d}$  can depend on the current parameters. Multiple algorithms exist that can solve a linear system of equations with linear constraints. Among them are methods using singular value decomposition, the direct elimination method, the Lagrangian multiplier method, and the null space method.[78]

An example of curve-fitting is presented to demonstrate the application of the Gauss-Newton method with and without constraints. A set of seven data points  $\{(x_i, y_i)\}$  is to be fitted with a function of the form

$$f(x, p_1, p_2) = \frac{p_1 x}{p_2 + x}. \quad (2.5)$$

The residual  $r_{k,i}$  of point  $i$  at iteration  $k$  is the distance between the data point and the fit function with the current values of  $a$  and  $b$

$$r_{k,i} = f(x_i, p_{k,1}, p_{k,2}) - y_i. \quad (2.6)$$

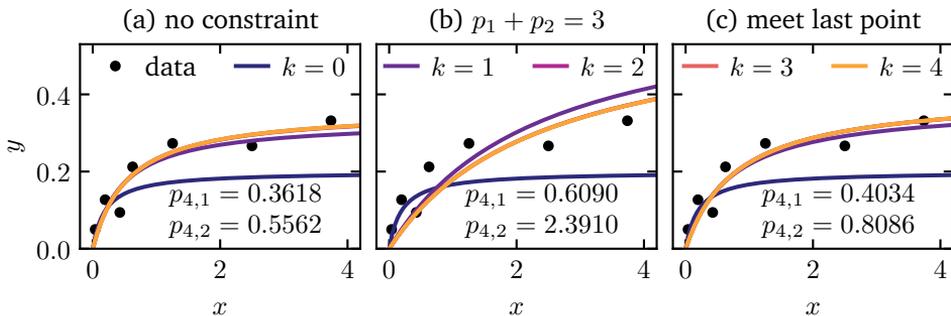


Figure 2.2: Gauss-Newton applied to fit data points with a curve of form  $f(x) = \frac{p_1 x}{p_2 + x}$ . The starting parameters are  $p_{0,1} = p_{0,2} = 0.2$ . In the middle and right panel, the constraint indicated in the panel title was considered to the update. At the bottom of each plot, the parameters after five iterations are given.

The Jacobian is a  $7 \times 2$  matrix with elements

$$J_{i,j} = \frac{\partial r_i}{\partial p_j} = \frac{\partial f(x_i, p_1, p_2)}{\partial p_j}. \quad (2.7)$$

The results for four iterations unconstrained Gauss-Newton method are shown in fig. 2.2 (a). For the given problem with two parameters, the Gauss-Newton method converges quickly, within two steps.

Next, two constraints to this fitting problem are considered. The first constraint allows only updates which lead to the sum of the two parameters to be three. This translates the following equation

$$p_{k+1,1} + p_{k+1,2} = p_{k,1} + \Delta p_{k,1} + p_{k,2} + \Delta p_{k,2} = 3. \quad (2.8)$$

Rewriting eq. (2.8) in the form of eq. (2.4) yields

$$\mathbf{K} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \mathbf{d}_k = (3 - (p_{k,1} + p_{k,2})). \quad (2.9)$$

In this case,  $\mathbf{d}$  changes with the iteration and is therefore written  $\mathbf{d}_k$ , while  $\mathbf{K}$  does not. The results are shown in fig. 2.2 (b). The parameters at all iterations except the initial guess add up to 3 (only the last values are shown).

As a second constraint example, the last data point  $(x_7, y_7)$  shall be always right on the curve. In contrast to the other constraint, a linearization is necessary because this constraint can not be expressed in terms of the parameters alone

$$f(x_7, p_{k+1,1}, p_{k+1,2}) = f(x_7, p_{k,1}, p_{k,2}) + \mathbf{J}_{k,7} \Delta \mathbf{p}_k + \mathcal{O}((\Delta \mathbf{p}_k)^2) = y_7. \quad (2.10)$$

---

Here,  $\mathbf{J}_{k,7}$  is the seventh column of the Jacobian at iteration  $k$ . Equation (2.10) can be rewritten in the form of eq. (2.4) with second and higher order terms neglected

$$\mathbf{K}_k = (\mathbf{J}_{k,7}) \quad \mathbf{d}_k = (y_7 - f(x_7, p_{k,1}, p_{k,2})). \quad (2.11)$$

The results of four iterations with this constraint are shown in fig. 2.2 (c). After the first iteration, the fit function does not yet meet the last data point. This inaccuracy is due to the linearization, which becomes less of an approximation the smaller  $\Delta p_k$  becomes. After the second iteration, the fit function meets the last data point, and the method converges. Some constraints, for example the first one in this example, can be considered directly in constructing the fit function. But this is often not tractable and the formulation with constraints allows for a lot of flexibility.

## 2.2 Bottom-up Coarse-Graining

This section gives a short overview of the theory behind bottom-up coarse-graining. First, the concept of mapping is introduced, and the *ideal* CG potential, the many-body potential of mean force, is derived. Then, the relative entropy is introduced as a measure for *non-ideal* CG potentials. Finally, two structure-based methods and one work-based method for coarse-graining are explained.

### 2.2.1 Mapping and the Many-Body Potential of Mean Force

The starting point of all coarse-graining methods is the mapping scheme. The mapping operator  $\mathcal{M}$  relates a fine-grained (FG) configuration  $\mathbf{r}$  to a coarse-grained (CG) configuration  $\mathbf{R}$

$$\mathbf{R} = \mathcal{M}(\mathbf{r}). \quad (2.12)$$

Here,  $\mathbf{r} = (x_1, y_1, z_1, \dots, x_n, y_n, z_n)$  and  $\mathbf{R} = (X_1, Y_1, Z_1, \dots, X_N, Y_N, Z_N)$  are vectors that contain all positions of the FG and CG particles, respectively. Usually, from the FG to the CG representation, the number of particles decreases, i.e.,  $n > N$ . In most cases, the mapping operator  $\mathcal{M}$  is linear and configuration-independent. Mapping can then be expressed as a multiplication with a matrix  $\mathbf{M}$  that contains the weights for the atoms' contribution to the bead position:  $\mathbf{R} = \mathbf{M} \mathbf{r}$ . A common scheme is mapping CG beads to the center of mass of a group of atoms. In some cases, the mapping operator depends on the configuration. An example would be grouping multiple water molecules into a single bead.[79] For that, the mapping has to be determined for every configuration by a clustering algorithm. In fig. 2.3 (a,b), an illustration of a fine-grained system and its mapped counterpart can be found.

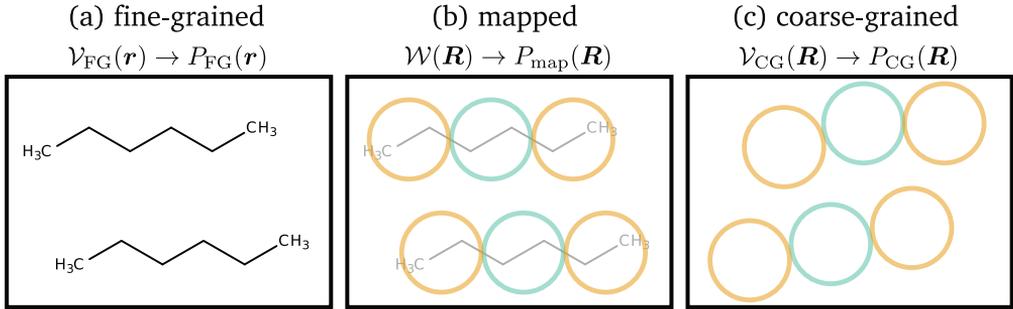


Figure 2.3: A fine-grained, mapped, and coarse-grained snapshot of two hexane molecules. The positions of the beads in the second panel are determined by the FG molecules, shown in light grey. The relation above each panel shows the potential function and the resulting probability density function.

In the canonical ensemble, the probability density function for a particular FG configuration  $\mathbf{r}$  is given by

$$P_{\text{FG}}(\mathbf{r}) = \frac{1}{Z_n} \exp(-\beta \mathcal{V}_{\text{FG}}(\mathbf{r})). \quad (2.13)$$

Here,  $\mathcal{V}_{\text{FG}}(\mathbf{r})$  is the potential energy as a function of a configuration  $\mathbf{r}$  and  $\beta = (k_{\text{B}}T)^{-1}$  is the thermodynamic beta where  $k_{\text{B}}$  is the Boltzmann constant and  $T$  is the temperature. The configuration integral  $Z_n$  of  $n$  particles is given by

$$Z_n = \int_{V^n} \exp(-\beta \mathcal{V}_{\text{FG}}(\mathbf{r})) \, d\mathbf{r} \quad (2.14)$$

with  $V$  being the volume of the system. The notation  $\int_{V^n}$  is short for  $n$  volume integrals over the system's volume with the coordinate vectors in  $\mathbf{r}$  as the integration variables.

To obtain the probability density function for a particular mapped configuration one needs to integrate over all possible FG configurations and filter with a Dirac delta function for FG configurations that map to the desired CG configuration[80]

$$P_{\text{map}}(\mathbf{R}) = \frac{\int_{V^n} P_{\text{FG}}(\mathbf{r}) \delta(\mathbf{R} - \mathcal{M}(\mathbf{r})) \, d\mathbf{r}}{\int_{V^N} \int_{V^n} P_{\text{FG}}(\mathbf{r}) \delta(\mathbf{R} - \mathcal{M}(\mathbf{r})) \, d\mathbf{r} \, d\mathbf{R}} \quad (2.15)$$

The Dirac delta function  $\delta(\mathbf{R})$  with an  $3N$ -dimensional arguments can be understood as a product of  $3N$  delta functions of the elements of  $\mathbf{R}$ . The denominator normalizes the probability density function. The mapped degrees of freedom (DOFs) naturally sample the same volume  $V$  as the FG DOFs. Again, the canonical ensemble is applied with the temperature  $T$  equal to the FG system's temperature.

The goal is to find an effective potential function  $\mathcal{W}(\mathbf{R})$  that generates the same probabilities through sampling in the canonical ensemble

$$P_{\text{map}}(\mathbf{R}) = P(\mathbf{R}) = \frac{\exp(-\beta\mathcal{W}(\mathbf{R}))}{\int_{V^N} \exp(-\beta\mathcal{W}(\mathbf{R})) d\mathbf{R}}. \quad (2.16)$$

Combining eqs. (2.13), (2.15) and (2.16), one finds for the effective potential the expression[81]

$$\mathcal{W}(\mathbf{R}) = -k_{\text{B}}T \log \left( \int_{V^n} \exp(-\beta\mathcal{V}_{\text{FG}}(\mathbf{r})) \delta(\mathbf{R} - \mathcal{M}(\mathbf{r})) d\mathbf{r} \right) + C \quad (2.17)$$

with  $\log()$  being the natural logarithm. The function  $\mathcal{W}(\mathbf{R})$  is called the **many-body potential of mean force** (m-PMF).[82] The constant  $C$  can be chosen freely, e.g., to normalize the m-PMF to yield zero for the separation of all molecules or make the argument of the logarithm in eq. (2.17) dimensionless. Note, that eq. (2.17) has the form of the relation between the free energy and probability, i.e.  $A = -k_{\text{B}}T \log(P)$ , and is sometimes called a *conditional* free energy.[83] Note also, that the m-PMF is a non-linear function of the temperature. The name m-PMF relates to the fact that its negative gradient  $-\nabla_{\mathbf{R}}\mathcal{W}(\mathbf{R})$  is the mean force averaged over FG forces. The m-PMF is the ideal CG potential for a given mapping operator, as it generates, by definition, the same configurations obtained from mapping the FG distributions. However, the direct application of eq. (2.17) is impractical for two reasons: (i) the resulting potential function is of high dimensionality and therefore impractical to store or compute, and (ii) the complete evaluation of the integral in eq. (2.17) is computationally prohibited for all but the most simple systems. Therefore, different methods have been established that aim at simpler objectives than the full m-PMF.

## 2.2.2 Relative Entropy

A general and systematic ansatz to the problem of finding the *best-possible* CG potential function  $\mathcal{V}_{\text{CG}}(\mathbf{R})$  is the relative entropy (RE) framework.[56, 84] For a given mapping operator, the relative entropy  $S_{\text{rel}}$  between a FG and a CG model is the Kullback-Leibler divergence between the probability density functions

$$S_{\text{rel}} = \int_{V^N} P_{\text{map}}(\mathbf{R}) \log \left( \frac{P_{\text{map}}(\mathbf{R})}{P_{\text{CG}}(\mathbf{R})} \right) d\mathbf{R} + \langle S_{\text{map}} \rangle_{\text{map}}. \quad (2.18)$$

The last term is the mapping entropy, which is determined by the mapping operator and independent of the CG potential. In the relative entropy context,  $P_{\text{map}}(\mathbf{R})$  is often written as  $P_{\text{FG}}(\mathbf{R})$  to display that the sampling of  $\mathbf{R}$  happens at the FG level. In the canonical ensemble, the definition of  $P_{\text{CG}}(\mathbf{R})$  is similar to eq. (2.16)

$$P_{\text{CG}}(\mathbf{R}) = P(\mathbf{R}) = \frac{1}{Z_N} \exp(-\beta\mathcal{V}_{\text{CG}}(\mathbf{R})) = \frac{\exp(-\beta\mathcal{V}_{\text{CG}}(\mathbf{R}))}{\int_{V^N} \exp(-\beta\mathcal{V}_{\text{CG}}(\mathbf{R})) d\mathbf{R}}. \quad (2.19)$$

Here, the configuration integral of the CG model  $Z_N$  was introduced. The relation between  $P_{\text{CG}}(\mathbf{R})$  and  $\mathcal{V}_{\text{CG}}(\mathbf{R})$  is depicted in fig. 2.3 (b). Since the aim is no longer to exactly match but to optimize the probability density function  $P_{\text{map}}(\mathbf{R})$ , the CG potential  $\mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})$  can be of any form, with a set of parameters  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$ . The relative entropy quantifies the overlap between the mapped reference probability density function and the CG model probability density function. Therefore, its minimization by varying the parameters of  $\mathcal{V}_{\text{CG}}$  offers a systematic route to an optimized CG potential.

If the relative entropy is minimal, its derivative with respect to all potential parameters has to be zero

$$\frac{\partial S_{\text{rel}}}{\partial \lambda_l} = 0 \quad \text{for all } l. \quad (2.20)$$

In the relative entropy definition (eq. (2.18)), only  $P_{\text{CG}}(\mathbf{R})$  (eq. (2.19)) depends on  $\boldsymbol{\lambda}$ . Applying the logarithmic derivative one obtains

$$\frac{\partial S_{\text{rel}}}{\partial \lambda_l} = \int_{V^N} -P_{\text{map}}(\mathbf{R}) \frac{1}{P_{\text{CG}}(\mathbf{R})} \frac{\partial P_{\text{CG}}(\mathbf{R})}{\partial \lambda_l} d\mathbf{R}. \quad (2.21)$$

The derivative of  $P_{\text{CG}}(\mathbf{R})$ , as defined in eq. (2.19), is acquired by the quotient and chain rule of differentiation

$$\frac{\partial P_{\text{CG}}(\mathbf{R})}{\partial \lambda_l} = \frac{-\beta \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} \exp(-\beta \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})) Z_N - \exp(-\beta \mathcal{V}_{\text{CG}}(\mathbf{R})) \frac{\partial Z_N}{\partial \lambda_l}}{Z_N^2} \quad (2.22)$$

with

$$\frac{\partial Z_N}{\partial \lambda_l} = \int_{V^N} -\beta \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} \exp(-\beta \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})) d\mathbf{R}. \quad (2.23)$$

By combining eqs. (2.20) to (2.23) one reaches

$$\begin{aligned} \frac{\partial S_{\text{rel}}}{\partial \lambda_l} &= \int_{V^N} \beta P_{\text{map}}(\mathbf{R}) \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} d\mathbf{R} - \int_{V^N} \beta P_{\text{CG}}(\mathbf{R}) \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} d\mathbf{R} \\ &\xrightarrow{\frac{\partial S_{\text{rel}}}{\partial \lambda_l} = 0} \left\langle \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} \right\rangle_{\text{map}} = \left\langle \frac{\partial \mathcal{V}_{\text{CG}}(\mathbf{R}, \boldsymbol{\lambda})}{\partial \lambda_l} \right\rangle_{\text{CG}}. \end{aligned} \quad (2.24)$$

This equation, while not immediately intuitive, is general and can be used in multiple ways. For example, it can be shown that coarse-graining a single component liquid to a Lennard-Jones pair potential, its optimal parameters are those where potential energy and virial pressure are matched.[56]

As mentioned before, a typical choice for the CG potential is a sum of pair interactions. For the following, a system containing only one type of CG beads is assumed; therefore, only a single pair potential function  $u(r)$  is needed. Note, a lowercase  $r$  is used to denote

the pair distance, even though it is a distance in between coarsened particles. For the potential energy sampled with  $P(\mathbf{R}_{\text{CG}})$ , one obtains

$$\langle \mathcal{V}_{\text{CG}}(\mathbf{R}) \rangle_{\text{CG}} = \frac{1}{2} \left\langle \sum_i^N \sum_j^N u(|\vec{R}_{ij}|) \right\rangle_{\text{CG}}. \quad (2.25)$$

The indices  $i$  and  $j$  are particle indices and  $\vec{R}_{ij} = (X_j - X_i, Y_j - Y_i, Z_j - Z_i)$  is the vector from particle  $i$  to particle  $j$ . Intramolecular contributions to the potential energy are ignored in eq. (2.25). The term  $\langle \mathcal{V}_{\text{CG}}(\mathbf{R}) \rangle_{\text{map}}$  is obtained similarly with the sampling in  $\langle \rangle_{\text{map}}$ . Equation (2.25) can be rewritten as

$$\langle \mathcal{V}_{\text{CG}}(\mathbf{R}) \rangle_{\text{CG}} = \frac{\rho N}{2} \int_0^\infty g_{\text{CG}}(r) u(r) 4\pi r^2 dr. \quad (2.26)$$

with the radial distribution function (RDF) defined as

$$g_{\text{CG}}(r) = \frac{1}{4\pi\rho N r^2} \left\langle \sum_i^N \sum_j^N \delta(r - |\vec{R}_{ij}|) \right\rangle_{\text{CG}}. \quad (2.27)$$

Here  $\rho$  is the number density  $N/V$ . The RDF of the mapped configurations  $g_{\text{map}}$  has the same definition but with the sampling in  $\langle \rangle_{\text{map}}$ . Upon inserting eq. (2.26) in eq. (2.24) and taking the functional derivative of the potential energy with respect to the pair potential at distance  $r$ , one reaches

$$\begin{aligned} \left\langle \frac{\delta \mathcal{V}_{\text{CG}}(\mathbf{R})}{\delta u(r)} \right\rangle_{\text{map}} &= \left\langle \frac{\delta \mathcal{V}_{\text{CG}}(\mathbf{R})}{\delta u(r)} \right\rangle_{\text{CG}} \\ \Rightarrow 2\pi\rho N r^2 g_{\text{map}}(r) &= 2\pi\rho N r^2 g_{\text{CG}}(r) \\ \Rightarrow g_{\text{map}}(r) &= g_{\text{CG}}(r) \end{aligned} \quad (2.28)$$

Accordingly, the relative entropy between an FG system and a CG system modeled by free-form pair potentials is minimal if the RDF obtained from the mapped coordinates and the CG model are the same.[56] This result generalizes to systems with multiple bead types.

The match of the RDFs when optimizing free-form pair potentials is only one result of the RE method. It has been used in combination with very different potential forms.[85, 86] For the minimization of  $S_{\text{rel}}$ , a Hessian matrix is obtained by double-differentiating eq. (2.18), and the Newton-Raphson method can be used to optimize the potential parameters.

### 2.2.3 Iterative Boltzmann Inversion and Inverse Monte Carlo

Historically, the motivation for structural coarse-graining came before the idea of RE coarse-graining. The initial paper on RE by M. Scott Shell was published in 2008, while methods like inverse Monte Carlo (IMC) and iterative Boltzmann inversion date back to 1995 and 1996.[50–52, 56]

Both IMC and IBI aim to match a target RDF iteratively by repeating four steps:

1. Using the current CG pair potential to run an MD simulation.
2. Calculating the RDF from the MD trajectory.
3. If the RDF is close enough to the target RDF, or the maximum number of iterations is reached, exit the iterations. Otherwise, go on.
4. Calculating a potential update (details depend on the method) and go back to 1).

Herein, the target RDF is  $g_{\text{map}}(r)$ , in the context of structural coarse-graining often called  $g_{\text{tgt}}$ . The iterative approach is necessary because matching the target RDF is an inverse problem and no closed-form expression exists to obtain the exact pair potential needed directly.

Before the iterative procedure, an initial pair potential guess  $u_0(r)$  is needed. The most common choice is the Boltzmann inversion, which takes the potential of mean force (PMF)  $w_{\text{map}}(r)$  as starting point

$$u_0(r) = w_{\text{map}}(r) = -k_{\text{B}}T \log g_{\text{map}}(r). \quad (2.29)$$

The index 0 indicates that this happens before the first iterations. Equation (2.29) is a weak approximation for dense fluids and sometimes called the low density approximation. With adapted normalization, eq. (2.29) applies to bond and angle potentials between CG beads[87], i.e.

$$u_{\text{bond},0}(r) = -k_{\text{B}}T \log P_{\text{bond}}(r) = -k_{\text{B}}T \log \frac{H_{\text{bond}}(r)}{4\pi r^2} \quad (2.30)$$

$$u_{\text{angle},0}(\theta) = -k_{\text{B}}T \log P_{\text{angle}}(\theta) = -k_{\text{B}}T \log \frac{H_{\text{angle}}(\theta)}{\sin \theta}. \quad (2.31)$$

Here,  $H(x)$  represents the histogram of sampled bond distances and angles from a trajectory. Equations (2.29) to (2.31) are applied to all non-bonded interactions, bonds, and angles whereby complete separability between the different DOFs is assumed.

The IBI method updates the pair potential iteratively using[51, 52]

$$\begin{aligned} u_{k+1}(r) &= u_k(r) + \Delta u_k(r) \\ &= u_k(r) + (w_{\text{map}}(r) - w_k(r)) \\ &= u_k(r) + k_{\text{B}}T \log \left( \frac{g_k(r)}{g_{\text{map}}(r)} \right) \end{aligned} \quad (2.32)$$

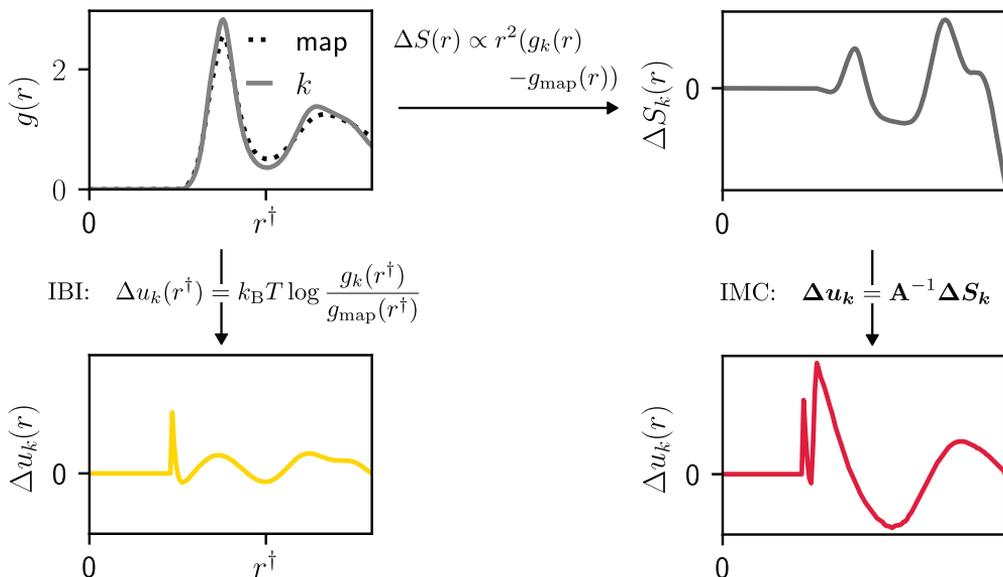


Figure 2.4: The process of obtaining an potential update  $\Delta u_k$  at iteration  $k$  with IBI and IMC. For IBI, the update is computed locally for each pair distance, illustrated for  $r^\dagger$ . With IMC, the potential update at each distance depends on  $\Delta S_k$  at all distances because the IMC matrix  $\mathbf{A}$  is not diagonal.

where the index  $k$  denotes the iteration and  $w_k(r)$  is related to  $g_k$  via eq. (2.29), respectively. The idea behind eq. (2.32) is a perturbation to a given potential that corrects for the mismatch between  $g_k$  and  $g_{\text{map}}$ . The magnitude of the perturbation is estimated by the difference in the PMF, which approximately predicts the perturbation necessary.[51] On the left side of fig. 2.4 the IBI update is depicted. Again, the pair potential and RDF are written as functions, while in practice both are evaluated on a grid. Equation (2.32) can also be used for intramolecular DOFs. In fact, IBI works well for intramolecular DOFs, as for them eq. (2.29) is a good approximation and they are often fairly decoupled.[88] Often, however, the bond and angle potentials are not further improved after their initialization with eqs. (2.30) and (2.31).[89, 90]

The IMC method is derived by expressing the average potential energy from an discrete set of parameters.[50] Starting from eq. (2.26) and discretizing the pair distance

$r$  to a regular grid with grid indices  $l$

$$\begin{aligned}\langle \mathcal{V}_{\text{CG}}(\mathbf{R}) \rangle_{\text{CG}} &= \frac{\rho N}{2} \int_0^\infty g_{\text{CG}}(r) u(r) 4\pi r^2 \, dr \\ &= \int_0^\infty S(r) u(r) \, dr \approx \sum_l S_l u_l.\end{aligned}\quad (2.33)$$

The pair potential values  $u_l$  take the role of  $\lambda_l$  in section 2.2.2. The distribution  $S(r)$  counts the number of pair interactions in a system at distance  $r$ . It is connected to the RDF  $g(r)$  by

$$S(r) = 2\pi\rho N r^2 g(r) = \frac{1}{2} \left\langle \sum_i^N \sum_j^N \delta(r - |\vec{R}_{ij}|) \right\rangle. \quad (2.34)$$

The discretization from  $S(r)$  and  $u(r)$  to  $S_l$  and  $u_l$ , respectively, works by integrating over a bin of width  $\Delta_r$  (the grid spacing) centered at  $r_l$

$$S_l = \int_{r_l - \frac{\Delta_r}{2}}^{r_l + \frac{\Delta_r}{2}} S(r) \, dr \quad \text{and} \quad u_l = \frac{1}{\Delta_r} \int_{r_l - \frac{\Delta_r}{2}}^{r_l + \frac{\Delta_r}{2}} u(r) \, dr. \quad (2.35)$$

To obtain the IMC scheme, the change of  $S_l$  with changes in the pair potential  $u_m$  is linearized

$$\Delta S_l = \sum_m \frac{\partial S_l}{\partial u_m} \Delta u_m + \mathcal{O}((\Delta u_m)^2) \quad \text{or} \quad \Delta \mathbf{S} = \mathbf{A} \Delta \mathbf{u} \quad (2.36)$$

with the elements of the Jacobian  $\mathbf{A}$ , called the IMC matrix, defined as

$$A_{lm} = \frac{\partial S_l}{\partial u_m}. \quad (2.37)$$

Equation (2.36) can be rearranged to obtain a potential update  $\Delta u_k = u_{k+1} - u_k$  from a structure mismatch  $\Delta S_k = S_k - S_{\text{map}}$ . Such an update, utilizing the inverse Jacobian, corresponds to a step in Newton's method, see eq. (2.2). In fig. 2.4, the IMC potential update is shown on the right side and the procedure compared to the IBI update. To obtain the IMC matrix, eqs. (2.19) and (2.33) the sampling of  $S_l$  is written explicitly as

$$S_l = S_{\text{CG},l} = \int_{V^N} S_l(\mathbf{R}) P_{\text{CG}}(\mathbf{R}) \, d\mathbf{R} = \frac{\int_{V^N} S_l \exp(-\beta \sum_m S_m u_m) \, d\mathbf{R}}{\int_{V^N} \exp(-\beta \sum_m S_m u_m) \, d\mathbf{R}}. \quad (2.38)$$

From there, the derivative is evaluated using the quotient rule and some algebra

$$\frac{\partial S_l}{\partial u_m} = -\beta (\langle S_l S_m \rangle_{\text{CG}} - \langle S_l \rangle_{\text{CG}} \langle S_m \rangle_{\text{CG}}). \quad (2.39)$$

---

---

With that, the IMC matrix can be sampled from the configurations generated by the current pair potential. The sampling of the matrix makes IMC computationally more expensive per iteration than IBI. The RE method, when using Newton’s method for the optimization of pair potentials, is equivalent to IMC.[91, 92]

## 2.2.4 Conditional Reversible Work

One interpretation of eq. (2.17) is the evaluation of all FG configurations that map to a single CG configuration  $\mathbf{R}$ . The conditional reversible work (CRW) method builds on this idea.[47] Again, the aim is to express the CG model with pair potentials. The target property is the reversible work in the atomistic representation, i.e., approaching two CG beads should take the same work as approaching the center of mass of the atoms that are represented by those beads. A *conditional* work is needed for multiple beads in a molecule in order to separate their contribution. A thermodynamic cycle is used, as shown in fig. 2.5. The idea is to approach the two groups of atoms representing the CG beads, once with and once without their intermolecular interactions turned on. The CRW pair potential  $u^{\text{CRW}}$  is obtained as

$$u^{\text{CRW}}(r) = W(r) - W^{\text{excl}}(r) + u^{\text{CRW}}(\infty) \quad (2.40)$$

The term  $W(r)$  represents the work it takes to approach the two groups of atoms to distance  $r$ . Similarly,  $W^{\text{excl}}(r)$  represents the work for the same distance with the intramolecular interaction between the atoms in the groups turned off. The last term can be assumed as zero at large for large distances. Compared to the other methods considered, CRW obtains the CG potential from simulations at fine-grained resolution. In practice, the two groups of atoms (i.e., their center of mass) are kept at a fixed distance, and the net force on the groups is averaged in an MD or Monte Carlo simulation. The force can then be integrated over a grid of distances to yield the work. Sampling the work terms can occur in vacuum or liquid, where the latter helps obtain more realistic configurations.

## 2.3 Integral Equation Theory

The integral equation theory is a field of statistical mechanics that connects microscopic structure with inter-particle potentials in fluids and allows for predicting its thermodynamic properties. An exact derivation of the basics of integral equation theory is lengthy and often deals with series of integrals. For handling their quickly growing complexity, diagrammatic methods were used to represent series of integrals. Instead of repeating the derivation here, an intuitive and historical approach is taken to introduce the basic concepts.

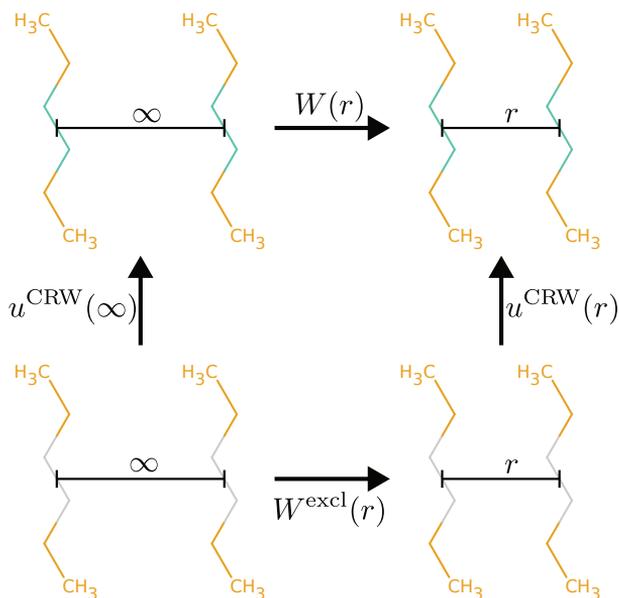


Figure 2.5: The thermodynamic cycle of the CRW method for determining the pair potential  $u^{\text{CRW}}(r)$  between the CG beads representing the two center methylene groups in hexane. Intermolecular interactions between the gray atoms are turned off in the bottom row. The CG beads mapped to the molecules are not explicitly shown, but the mapping is equivalent to fig. 2.3.

### 2.3.1 Ornstein-Zernike Equation

The basic ideas were developed in 1914 by Leonard S. Ornstein and Frits Zernike.[93] They found it impossible to develop a valid approximation for the density fluctuations in a subvolume of a system near the critical point under the assumption that the subvolumes are independent. Taking into account the correlations between subvolumes, they derived an equation of state that correctly predicts the divergence of the compressibility at the critical point. That is in contrast to the van der Waals equation which does not correctly predict this experimental results. The **Ornstein-Zernike (OZ) equation** for a system of identical particles can be written as

$$h(\vec{r}_{12}) = c(\vec{r}_{12}) + \int_V c(\vec{r}_{13})\rho(\vec{r}_3)h(\vec{r}_{23})d^3\vec{r}_3 \quad (2.41)$$

Here,  $h(\vec{r}) = g(\vec{r}) - 1$  is the total correlation function and  $g(\vec{r})$  is the RDF. The single particle density is denoted by  $\rho(\vec{r})$ . The notation  $\vec{r}_{12}$  indicates the vector between particles

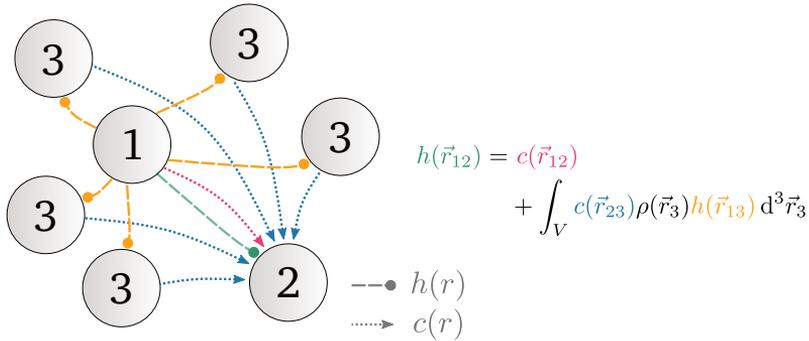


Figure 2.6: An *instantaneous* interpretation of the Ornstein-Zernike equation in a small system of atoms. Atom 1 is influencing atom 2 directly, shown as a dotted red arrow. Indirect influence occurs from atom 2 structuring some atoms with label 3 (dashed yellow), affecting atom 1 (dotted blue). The equation is slightly different from eq. (2.41) by using the commutativity of convolution. Note that the OZ equation is defined in terms of distribution functions. A more faithful but less accessible representation would contain volume elements with similar interactions.

1 and 2, i.e.  $\vec{r}_{12} = \vec{r}_2 - \vec{r}_1$ . The newly introduced function  $c(\vec{r})$  is the direct correlation function (DCF). It is defined by the OZ equation and can be thought of as the direct influence that one particle inflicts on another particle. The DCF is also connected to the pair potential, a relation that will be elaborated soon. An illustration of the rationale behind eq. (2.41) is shown in fig. 2.6. An alternative definition of the DCF by the functional derivative of the excess part of the free-energy functional is possible but not given here.[94]

If the system is uniform and isotropic, eq. (2.41) can be simplified to

$$h(r) = c(r) + \rho \int_V c(r') h(|\vec{r}' - \vec{r}|) d^3\vec{r}' \quad \text{or} \quad h(r) = c(r) + \rho c(r) * h(r) \quad (2.42)$$

Here,  $r$  is simply the pair distance (note: in section 2.2 the same symbol is used for the vector of all particle positions). The operator  $*$  is the radially symmetric convolution over the volume of the system. An interesting property of the OZ equation is its recursive nature. Upon repeatedly reinserting  $h$  in the right side of eq. (2.42) one obtains

$$h(r) = c(r) + \rho c(r) * c(r) + \rho^2 c(r) * c(r) * c(r) + \dots \quad (2.43)$$

It demonstrates how a liquid's RDF results from direct and a large number of indirect interactions, propagated via an increasing number of mediating particles.[94]

The OZ equation was first defined for pure fluids, but has equivalent form for mixtures

$$h_{\alpha\beta}(r) = c_{\alpha\beta}(r) + \sum_{\gamma} \rho_{\gamma} \int_V c_{\alpha\gamma}(r') h_{\beta\gamma}(|\vec{r}' - \vec{r}|) d^3\vec{r}'$$

or  $\hat{\mathbf{h}}(\nu) = \hat{\mathbf{c}}(\nu) + \hat{\mathbf{c}}(\nu)\rho\hat{\mathbf{h}}(\nu).$  (2.44)

The matrix  $\rho$  is diagonal with elements  $(\rho)_{\gamma\gamma} = \rho_{\gamma}$ . The matrices  $\mathbf{h}$  and  $\mathbf{c}$  are symmetric with elements  $(\mathbf{h})_{\alpha\beta} = (\mathbf{h})_{\beta\alpha} = h_{\alpha\beta}$  with the indices  $\alpha, \beta, \gamma$  symbolize particle types. The hat  $\hat{\phantom{x}}$  denotes the radially symmetric 3D Fourier transform. The Fourier transform simplifies the equation by converting convolution into multiplication. The matrix equation can thereby be solved independently for each wave vector  $\nu$ .

### 2.3.2 Closure Relations

The OZ equation alone is just a definition of the direct correlation function. A closure relation is needed that connects the DCF with the pair potential  $u(r)$ . Such a relation is of great benefit since it results in a closed set of equations. It allows the calculation of the RDF from a pair potential for different densities and temperatures. From there, three routes can be used to calculate, e.g., the pressure:

1. calculating the potential energy from the integral over pair potential and RDF and taking its derivative with respect to volume,
2. calculating the compressibility from the RDF and integrating over the volume,
3. integrating over pair force and RDF to obtain the pressure directly.

The knowledge about microscopic interactions is thereby used to obtain a macroscopic equation of state.

Three well-known closure relations are the hypernetted-chain (HNC) closure, Percus-Yevick (PY) closure, and the mean spherical approximation (MSA)[95–97]

$$c(r) = h(r) - \frac{u(r)}{k_{\text{B}}T} - \log g(r), \quad \text{HNC (2.45)}$$

$$c(r) = g(r) \left( 1 - e^{\frac{u(r)}{k_{\text{B}}T}} \right), \quad \text{PY (2.46)}$$

$$c(r) = -\frac{u(r)}{k_{\text{B}}T} \quad \text{if } r > \sigma_{\text{HS}}. \quad \text{MSA (2.47)}$$

The MSA is only defined for potentials with a hard-sphere repulsion with  $u(r) = \infty$  if  $r < \sigma_{\text{HS}}$ . The closure relations can be motivated in different ways, which are omitted here for brevity.[94] It shall be noted that many more closure relations exist, and all are approximations. One problem that arises from the approximate nature of the closure relations is that the routes mentioned above would not all yield the same pressure.[98]

The three closure relations above share some properties. If the density approaches zero, the OZ equation predicts  $h(r) = c(r)$ . Inserting into eqs. (2.45) to (2.47), one obtains the relations

$$\lim_{\rho \rightarrow 0} g(r) = e^{-\frac{u(r)}{k_B T}} \quad \text{HNC, PY (2.48)}$$

$$\lim_{\rho \rightarrow 0} g(r) = 1 - \frac{u(r)}{k_B T} \approx e^{-\frac{u(r)}{k_B T}}. \quad \text{MSA (2.49)}$$

This equation is essentially the Boltzmann distribution law. The limit shows that the OZ and closure equations can be understood as extensions to the low-density approximation. Another common property among closure relations is the trend of the DCF at large distances. The RDF quickly approaches one, and the potential becomes small. Under these conditions, one obtains from all three closure relations

$$\lim_{r \rightarrow \infty} \lim_{g \rightarrow 1} c(r) = \frac{-u(r)}{k_B T}. \quad (2.50)$$

This asymptotic behavior is the expected result that can be independently obtained from perturbation theory.[94] It was the motivation for the mathematical form of the MSA.

Since the late 1950s, integral equations have successfully predicted the properties of particles interacting with hard-sphere (HS) and square well (SW) potentials.[96, 99] Also, the Lennard-Jones (LJ) has been subject to early studies.[100] At the same time, Monte-Carlo simulations became computationally feasible.[101] They would serve as a reference in most works to compare with results from integral equation theory. An analytic solution for the pure and mixed HS fluids using the PY closure was found.[102–104] The packing fraction fully describes a pure liquid modeled by an HS model. In contrast, the SW model introduces an attractive well and makes the structure temperature-dependent. It was, therefore, of high interest as a model for actual fluids. The PY equation and MSA were applied to the SW potential to obtain a liquid equation of states.[99, 105]

### 2.3.3 Reference Interaction Site Model

An important conceptual step forward was the development of the reference interaction site model (RISM).[106] A simple argument for its derivation, made by Wim Briels[107], shall be outlined here. Starting point is eq. (2.44) and assuming all distributions to be sum of intra- and intermolecular terms (superscript full)

$$\hat{\mathbf{h}}^{\text{full}} = \hat{\mathbf{c}}^{\text{full}} \left( 1 + \rho \hat{\mathbf{h}}^{\text{full}} \right). \quad (2.51)$$

The number density matrix  $\rho$  is diagonal with diagonal elements  $\rho_\alpha$ , the densities of the atoms. The total correlation function is split in two parts, one of which is the usual

intermolecular TCF  $\hat{h}$

$$\hat{h}^{\text{full}} \rho = \hat{h}^{\text{intra}} \rho + \hat{h} \rho = (\hat{\omega} - 1) + \hat{h} \rho. \quad (2.52)$$

The distribution  $\omega_{\alpha\beta}(r)$  describes the probability to find atoms  $\alpha$  and  $\beta$  at distance  $r$  inside a molecule. For  $\alpha = \beta$  the distribution  $\omega_{\alpha\beta}(r)$  is  $\delta(r)$ , the Dirac Delta distribution. The direct correlation function also needs to be split in two parts. One inserts the intramolecular part from eq. (2.52) into an assumed OZ equation times  $\rho$  that holds only for the intramolecular terms with the same form as eq. (2.51) to obtain the intramolecular DCF

$$\begin{aligned} \hat{h}^{\text{intra}} \rho &= \hat{c}^{\text{intra}} (\rho + \rho \hat{h}^{\text{intra}} \rho) \\ (\hat{\omega} - 1) &= \hat{c}^{\text{intra}} (\rho + \rho(\hat{\omega} - 1)) \\ \hat{c}^{\text{intra}} \rho &= 1 - \hat{\omega}^{-1} \end{aligned} \quad (2.53)$$

It is assumed that the DCF can be split into two parts, the former of which was just obtained

$$\hat{c}^{\text{full}} \rho = \hat{c}^{\text{intra}} \rho + \hat{c} \rho = (1 - \hat{\omega}^{-1}) + \hat{c} \rho. \quad (2.54)$$

Here,  $\hat{c}$  is the intermolecular DCF. Inserting eq. (2.52) and eq. (2.53) in eq. (2.51) one obtains

$$\begin{aligned} \hat{h}^{\text{full}} \rho &= \hat{c}^{\text{full}} \rho (1 + \hat{h}^{\text{full}} \rho) \\ \hat{\omega} - 1 + \hat{h} \rho &= (1 - \hat{\omega}^{-1} + \hat{c} \rho) (\hat{\omega} + \hat{h} \rho) \\ \hat{h} \rho &= \hat{c} \rho \hat{\omega} + \hat{h} \rho - \hat{\omega}^{-1} \hat{h} \rho + \hat{c} \rho \hat{h} \rho \\ \hat{h} &= \hat{\omega} \hat{c} \hat{\omega} + \hat{\omega} \hat{c} \rho \hat{h} \end{aligned} \quad (2.55)$$

This is the RISM equation. The fact that  $\hat{\omega}$  and  $\rho$  commute was used in the derivation. They commute because  $\hat{\omega}$  has a block-diagonal structure and  $\rho$  is diagonal and has equal diagonal elements for each block in  $\hat{\omega}$  since all atoms in a molecule have the same number density. A variant of RISM that has found use in biochemistry is the three-dimensional generalization of the reference interaction site model (3D-RISM).[108, 109] It can be used to calculate the hydration structure and solvation free energy of solvents from small molecules to proteins.[110] The 3D-RISM method is not further discussed here since it was not used in this work.

Equation (2.55) is the RISM equation and an illustration thereof is provided in fig. 2.7. It can be seen as a definition of the DCF, similar to eq. (2.44) for atomic systems, but it contains approximations that the OZ equation does not contain. In eq. (2.53), by using the OZ equation only for intramolecular terms, it is assumed that the intramolecular DCF fully describes the intramolecular structure. However, this is only true for small stiff molecules,

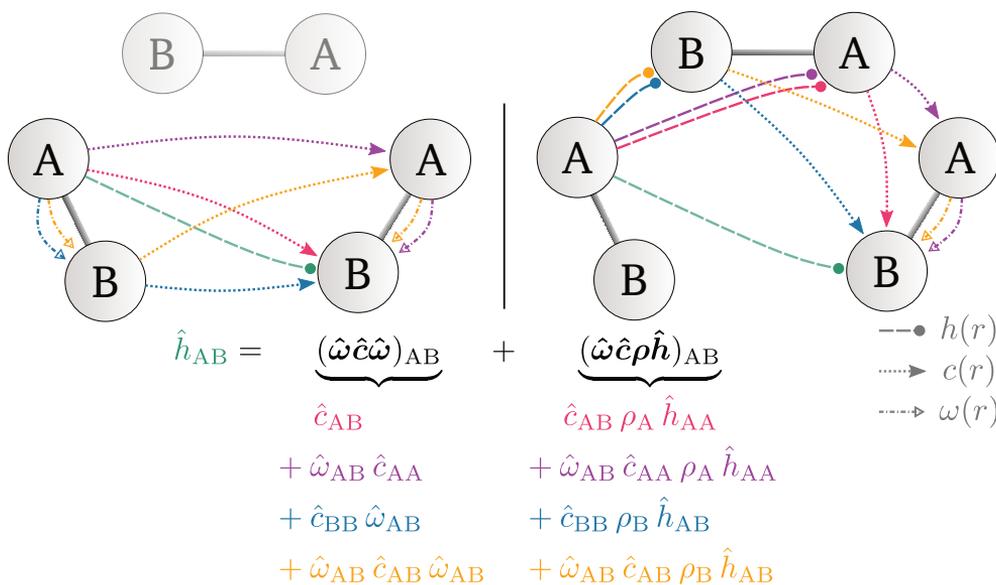


Figure 2.7: An interpretation (similar to fig. 2.6) of the RISM equation for a system with molecules of two atoms. The first term represents all interactions that influence  $h_{AB}$  and occur directly between two molecules, shown in the left diagram. The right diagram represents the second term with all indirect interactions over other molecules in the system (only one shown). For brevity, Fourier transformed distribution functions are used in the equation.

whereas “softer” molecules’ internal DOFs are influenced by intermolecular interactions. Another approximation comes from describing the intramolecular structure only by a set of one-dimensional distribution functions. This inexactness becomes a problem for larger molecules with distinct internal structure. A somewhat exotic but educational example is the difference between a pure enantiomer and a racemic mixture of chiral molecules. All atom-atom potentials and intramolecular distances are equal in both cases. Therefore, the RISM equation (with some closure relation) would predict equal structure and properties for the two liquids, contrary to the expected differences.[111] David Chandler has noted that higher-order intramolecular distributions would be needed to improve the RISM ansatz.[112] The Chandler–Silbey–Ladanyi equation has been developed in that direction but comes at the cost of increased numerical complexity.[113]

For melts of linear polymers, the RISM equation can be used with certain approximations to the intramolecular structure.[114] It is assumed that end-group-effects are negligible, which is only exact for ring polymers. Then, because there is equivalence of

---

---

all matrix elements  $h_{\alpha\beta}$  and  $c_{\alpha\beta}$ , the matrix eq. (2.55) reduces to a single equation

$$\hat{c} = \hat{w}\hat{c}\hat{w} + \rho_m\hat{w}\hat{c}\hat{h}. \quad (2.56)$$

Here,  $\rho_m$  is the number density of the monomer and  $w(r) = \frac{1}{n} \sum_{\alpha=1}^n \sum_{\beta=1}^n \omega_{\alpha\beta}(r)$ , where  $n$  is the number of monomers. Equation (2.56) is the basic idea of polymer-RISM or PRISM. Often,  $w(r)$  is obtained from some idealized model, such as the freely-jointed chain model. PRISM has been adapted further to work for binary blends and solvated polymers.[115, 116]

---

---

## 3 Iterative integral equation methods for structural coarse-graining

---

### Abstract

In this paper, new Newton and Gauss-Newton methods for iterative coarse-graining based on integral equation theory are evaluated and extended. In these methods, the potential update is calculated from the current and target radial distribution function, similar to iterative Boltzmann inversion, but gives a potential update of quality comparable with inverse Monte-Carlo. This works well for the coarse-graining of molecules to single beads which we demonstrate for water. We also extend the methods to systems that include coarse-grained bonded interactions and examine their convergence behavior. Finally, using the Gauss-Newton method with constraints, we derive a model for single bead methanol in implicit water which matches the osmotic pressure of the atomistic reference. An implementation of all new methods is provided for the open-source VOTCA package.

### 3.1 Introduction

Structure-based coarse-graining aims at representing structural information of a fine-grained system through a coarse-grained (CG) model with fewer degrees of freedom. The goal of structure-based coarse-graining is the approximation of a CG N-body potential of mean force with pair potentials, that reproduces a set of distribution functions.[117] This can be thought of as an optimization problem: Optimize a potential until it reproduces the radial distribution function (RDF). For isotropic single-particle systems, the Henderson uniqueness theorem states that there is a bijective map between pair potential and RDF, which suggests that an optimal potential exists.[48, 49] The two most common structure-based coarse-graining methods are Iterative Boltzmann inversion (IBI)[52] and Inverse Monte Carlo (IMC)[50]. IBI and IMC are applied widely in systems such as ionic liquids, polymers, and biological systems, where they help modelling time and length scales that are too expensive with atomistic MD.[38, 118–121] Both methods iteratively improve pair potentials based on the distance from the current to the target distribution.[54] Delbary et al. have proposed two new Newton-type schemes, HNCN and IHNC, and

the Gauss-Newton scheme HNCGN which are based on integral equation theory and conceptual compromises between IBI and IMC.[75]

They showed that their method can retrieve a Lennard-Jones (LJ) potential and generated a potential from experimental Argon data. In this work, we are comparing those new schemes with the IBI and IMC methods for coarse-graining molecular liquids.

The potential update of the Newton scheme is given by

$$u_{k+1} = u_k - \mathbf{J}^{-1} (g_k - g_{\text{tgt}}). \quad (3.1)$$

Here,  $u_k$  and  $g_k$  are the potential and RDF at iteration  $k$ , respectively, and  $\mathbf{J}$  is the Jacobian with elements  $J_{\alpha\gamma} = \frac{\partial g_\alpha}{\partial u_\gamma}$ . The analytical form of the map  $u(g)$  is unknown, and the usual connection is to calculate  $g$  from  $u$  by molecular dynamics (MD) or Monte Carlo (MC) simulations. The IBI potential update results from the connection of  $u$  and  $g$  in the low density approximation, i.e. the direct Boltzmann inversion  $g \approx e^{-\beta u}$  with  $\beta = \frac{1}{k_B T}$ . This makes the IBI update a rough approximation to the Newton scheme if applied to liquids, that sometimes needs hundreds of iterations for convergence.[54, 122] In IMC, the Jacobian is calculated from cross-correlations in the distance distribution in the system. With enough sampling the IMC Jacobian approximates the exact Jacobian, which is why the authors call their Ansatz ‘‘Newton inversion’’.[117] For simple systems IMC converges in under 10 iterations. Lack of convergence, which is typical for Newton methods far from the optimum, can be addressed using regularization.[123] Nonetheless, the convergence can be slow, in particular for systems with multiple components.[121, 124] This can make IMC coarse-graining computationally costly since the sampling of the IMC Jacobian needs long trajectories.

While direct Boltzmann inversion (BI) provides a reasonable estimate of the effective pair potential for low densities, this is no longer the case at liquid densities where direct BI leads to multibody contributions included in the effective pair potential. Here, the Ornstein-Zernike (OZ) equation combined with a hypernetted-chain (HNC) or Percus-Yevick (PY) closure provides a better approximation of the effective pair potential.[71] The HNCN method (hypernetted-chain Newton) uses the HNC closure to derive an approximation to the Jacobian, the input to the potential update scheme being the same as for IBI: the RDFs from the current potential.[75] This makes the computational cost per iteration to be comparable to IBI and potentially cheaper than IMC. At the same time, the number of iterations for convergence in single-component systems is comparable to IMC. IBI converges much slower for similar systems due to the crude approximation to the Jacobian.[53]

Using integral equations is not new in the field of molecular coarse-graining. This comes at no surprise as it provides an analytical connection between structure and potentials in liquid systems. Guenza et al. have established non-iterative methods for obtaining potentials for CG polymer melts.[69, 125] By approximating the intramolecular distribution function with an analytical function, it is possible to solve the PRISM equation.[126] Due to the analytical nature of the equations, transferability across different

---

---

resolutions or densities can be induced by fitting general trends in the direct correlation function.[70] Mashayak et al. have demonstrated that one can use integral equations for a good potential guess which can subsequently be improved by IBI.[73]. The method by Levesque et al. uses integral equations in a secant method way.[74, 127] It comes closest to the methods discussed in this paper but differs in that the Jacobian is never calculated.

Potentials derived from structural coarse-graining are state point dependent.[128] To improve transferability, multistate-IBI can be used to fit the structure of several state points with one common potential.[129] CG potentials also don't generally represent the thermodynamic properties of their reference system. IBI can be pressure corrected by adding a ramp potential after each iteration.[52, 61] Interestingly, we could not find any studies which implemented pressure ramps or constraints with IMC, even though a ramp correction is implemented in both common IMC codes, VOTCA, and Magic.[87, 130] Instead, an extension to constrain the surface tension during IMC was demonstrated.[123] A different approach is to extend the Hamiltonian with a term that only influences the pressure of the CG system, but not the structure, which can be extrapolated to different temperatures.[64, 67] For the description of multi-phase systems, this approach was developed to utilize local density potentials which are optimized through the relative entropy method.[131] For the relative entropy method itself it was demonstrated that through the use of Lagrange multipliers the pressure can be constrained during the potential optimization.[132] Another property that can be incorporated as a target is the Kirkwood-Buff integral, as demonstrated for IBI.[133, 134] IMC has been previously adapted to incorporate a constraint for the area compressibility of a phospholipid bilayer.[123] To apply constraints into the integral equation methods, Delbary et al. have reformulated their method to a Gauss-Newton method HNCGN.[75] It allows multiple constraints to be incorporated into the updating scheme.

This paper is organized as follows: In the theory section, we shortly recapitulate the basis of the HNCN and the HNCGN methods. We propose a new scheme derived from the Reference Interaction Site Model (RISM) to expand the methods to systems where the CG representation includes bonds. Some variants of the HNCN method and a scheme for RDF extrapolation are defined and examined. We apply IBI, IMC, HNCN, and HNCGN on water, hexane, and naphthalene, where the two latter systems are the test case for CG molecules with bonds. We examine in detail differences in the Jacobians and their physical meaning. Those differences explain the variance in the convergence behavior of the different methods. Finally, we use the HNCGN method with constrained pressure for the coarse-graining of methanol in water to an implicit solvent system with correct osmotic pressure.

## 3.2 Theory

### 3.2.1 Newton Iteration

The derivation of the integral equation methods for coarse-graining atomistic systems is described with greater mathematical rigor in the original paper.[75] Here we will just point out the steps important for understanding the method and results. The radially symmetric OZ equation for a system with one particle type

$$h = c + \rho h * c \quad (3.2)$$

and the HNC closure

$$u = \frac{1}{\beta} (h - \log(g) - c). \quad (3.3)$$

can be used to express the potential as a function of  $h$ . [94] Here,  $h = g - 1$  is the total correlation function,  $c$  is the direct correlation function,  $\rho$  is the particle number density,  $\beta$  is  $(k_B T)^{-1}$ , and  $\log$  the natural logarithm. All variables except  $\rho$  and  $\beta$  are functions of the particle-particle distance  $r$ . The operator  $*$  denotes a three-dimensional convolution. To solve the OZ equation for  $c$  we switch to Fourier space where the convolution becomes a multiplication. The Fourier and inverse Fourier transform over  $r$  in 3D for a radially symmetric function  $f$  are defined as

$$\hat{f}(\omega) = \mathcal{F}(f) = \int_{\mathbb{R}^3} f(|\mathbf{r}|) e^{-i\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} = \frac{2}{\omega} \int_0^\infty f(r) r \sin(2\pi r \omega) dr \quad (3.4)$$

$$f(r) = \mathcal{F}^{-1}(\hat{f}) = \frac{2}{r} \int_0^\infty \hat{f}(\omega) \omega \sin(2\pi r \omega) d\omega. \quad (3.5)$$

The resulting expression for  $\hat{c}$  from equation (3.2) is

$$\hat{c} = \frac{\hat{h}}{1 + \rho \hat{h}}. \quad (3.6)$$

It is transformed back to real space and when inserted in equation (3.3) gives an expression for the potential as a functional of the RDF

$$u = \frac{1}{\beta} (h - \log(g) - \mathcal{F}^{-1}(\hat{c})) = \frac{1}{\beta} \left( h - \log(g) - \mathcal{F}^{-1} \left( \frac{\hat{h}}{1 + \rho \hat{h}} \right) \right). \quad (3.7)$$

For the Newton or Gauss-Newton type potential update we need the corresponding approximation of the inverse of the Jacobian, which is the derivative of  $u$  by  $g$ . From the HNC closure (equation (3.3)) we obtain

$$\frac{du}{dg} = \frac{1}{\beta} \left( 1 - \frac{1}{g} - \frac{dc}{dg} \right). \quad (3.8)$$

Since  $c$  is an operator on  $g$ , so is the derivative  $\frac{dc}{dg}$ , which can be calculated in Fourier space

$$\frac{dc}{dg} = \mathcal{F}^{-1} \left( \frac{d\hat{c}}{d\hat{g}} \right) \mathcal{F} = \mathcal{F}^{-1} \left( \frac{1}{(1 + \rho\hat{h})^2} \right) \mathcal{F}. \quad (3.9)$$

For the discrete case this can be written as a matrix, where the Fourier operator becomes a Fourier matrix. The matrix in equation (3.9) produces the non-diagonal elements of the Jacobian's inverse in equation (3.1)

$$u_{k+1} = u_k - \frac{1}{\beta} \left[ \left( 1 - \frac{1}{g_k} \right) - \mathcal{F}^{-1} \left( \frac{1}{(1 + \rho\hat{h}_k)^2} \right) \mathcal{F} \right] (g_k - g_{\text{tgt}}). \quad (3.10)$$

This is called the hypernetted-chain Newton (HNCN) iteration. The term in the square brackets is the inverse of the Jacobian. It does not need to be explicitly calculated if one takes  $g_k - g_{\text{tgt}}$  into Fourier space. In the original paper, it is described that for the HNCN method the potential should be calculated on a longer range to obtain reasonable low-frequency values for  $\hat{h}$ . The tail of the potential update is then cut off. An alternative is to calculate the Jacobian explicitly from long RDFs and only use the square cutout that describes  $\frac{du_{\text{short}}}{dg_{\text{short}}}$ . It is then inverted again and equation (3.10) is applied on the short-range; this we call the HNCN Jacobian cutout (jc) method. The latter method applied to RDFs, where the tails have been extrapolated as will be explained in section 3.2.4 we call HNCN extrapolated (ex). The straightforward application of equation (3.10) with the RDF and the potential on the same, short-range we call HNCN short distribution (sd).

Equation (3.10) can be modified slightly by approximating  $-(g_k - g_{\text{tgt}})/g_k \approx \log(1 + (g_{\text{tgt}} - g_k)/g_k) = \log(g_{\text{tgt}}/g_k)$  such that the first term resembles the IBI update, which is then called the inverse hypernetted-chain (IHNC) iteration.[75]

$$u_{k+1} = u_k - \underbrace{\frac{1}{\beta} \log \left( \frac{g_{\text{tgt}}}{g_k} \right)}_{\Delta u_{\text{IBI}}} - \frac{1}{\beta} \left[ (g_k - g_{\text{tgt}}) - \mathcal{F}^{-1} \left( \frac{(\hat{g}_k - \hat{g}_{\text{tgt}})}{(1 + \rho\hat{h}_k)^2} \right) \right]. \quad (3.11)$$

### 3.2.2 Gauss-Newton Iteration

Delbary et al. introduced a Gauss-Newton type scheme, which has two advantages over a pure Newton scheme: (i) the calculated RDF can be longer ranged than the potential, which can naturally be reflected in a non-square Jacobian, and (ii) the scheme allows for the inclusion of one or multiple constraints.[75] Again, we only show the important parts of the scheme and refer to the original paper for details and mathematical rigor. We switch to a discretized notation with distributions and potentials becoming vectors and operators becoming matrices. The problem of finding the potential is reformulated

as a minimization

$$\arg \min_{\mathbf{w}_k} \|\mathbf{g}_{\text{tgt}} - \mathbf{g}_k - \mathbf{U}^{-1} \mathbf{A}_0 \mathbf{w}_k\|_2. \quad (3.12)$$

$\mathbf{U}^{-1} \mathbf{A}_0$  represents the Jacobian matrix and  $\mathbf{w}_k = \mathbf{f}_{k+1} - \mathbf{f}_k$  is the force update in iteration  $k$ . Writing the iteration in terms of the force allows for a pressure constraint later. The matrix  $\mathbf{A}_0$  is  $\Delta r$  times an upper unitriangular matrix stacked on a block of zeros. The unitriangular part acts as an antiderivative operator that transforms the force update to the potential update

$$\mathbf{u}_{k+1} - \mathbf{u}_k = \mathbf{A}_0 \mathbf{w}_k. \quad (3.13)$$

The block of zeros in  $\mathbf{A}_0$  makes it rectangular and causes the potential update to be zero beyond a desired cut-off, but based on the whole range of the input RDF. Matrix  $\mathbf{U}$  represents the Jacobian with respect to the potential which we already encountered in equation (3.10)

$$\mathbf{U}^{-1} = \frac{1}{\beta} \left( \text{diag} \left( 1 - \frac{1}{g_{\text{tgt}}} \right) - \mathbf{F}^{-1} \text{diag} \left( \frac{1}{(1 + \rho \hat{h}_{\text{tgt}})^2} \right) \mathbf{F} \right). \quad (3.14)$$

$\mathbf{F}$  is the Fourier matrix.

The Gauss-Newton scheme allows us to add constraints to the potential update. A classic constraint in structure-based coarse-graining is the pressure. The Henderson theorem would predict that only one potential can reproduce a given RDF. Previous studies showed, that certain changes to the potential have little effect on the distribution function.[61] This motivates the scheme, where the pressure is enforced and the RDF is matched as good as possible. The constraint, derived from the virial, has the form

$$\mathbf{l}^T \mathbf{w}_k = p_{\text{tgt}} - p_k, \quad (3.15)$$

where  $p_{\text{tgt}}$  and  $p_k$  are the target and current pressure, respectively. The elements of  $\mathbf{l}$  are given by

$$l_\alpha = \frac{2}{3} \pi \rho^2 \frac{g_{\text{tgt},\alpha} + g_{\text{tgt},\alpha+1}}{2} \frac{r_{\alpha+1}^4 - r_\alpha^4}{4} \quad (3.16)$$

The constraint is exact if  $g_{\text{tgt}} = g_{k+1}$  so the pressure will not be precisely met in early iterations.

### 3.2.3 Extension to Symmetric Molecules with Internal Bonds

Here we extend the methodology from sections 3.2.1 and 3.2.2 to one component systems where the CG molecules consist of  $n$  identical beads. The beads have to be bonded in a way that makes them indistinguishable. Starting point is the RISM-OZ relation[94]

$$\hat{\mathbf{h}} = \hat{\Omega} \hat{\mathbf{c}} \hat{\Omega} + \hat{\Omega} \hat{\mathbf{c}} \rho \hat{\mathbf{h}} = \hat{\Omega} \hat{\mathbf{c}} (\hat{\Omega} + \rho \hat{\mathbf{h}}). \quad (3.17)$$

where  $\hat{\mathbf{h}}$  and  $\hat{\mathbf{c}}$  are matrices with elements  $\hat{h}_{ij}$  and  $\hat{c}_{ij}$  of sites  $i$  and  $j$ , respectively. Each site in a molecule has a separate index, ignoring indistinguishability for now. Thereby each site has the same number density  $\rho_i = \rho$ .  $\Omega_{ij}$  represents the intramolecular distribution function between sites  $i$  and  $j$

$$\Omega_{ij} = \rho(G_{ij} - g_{ij}) + \delta_{ij}\delta(\mathbf{r}) \iff \hat{\Omega}_{ij} = \rho(\hat{G}_{ij} - \hat{g}_{ij}) + \delta_{ij}. \quad (3.18)$$

Its Fourier transform  $\hat{\Omega}_{ij}$  is the intramolecular structure factor. Note that the delta distribution of vector  $\mathbf{r}$  (not the same as  $\delta(r)$ ) becomes one by the 3D Fourier transform.  $G_{ij}$  is defined similar to a normal RDF, but in addition counts sites that are on the same molecule, except for the reference site. In the appendix we show how it can be expressed in terms of the average intramolecular distribution function under the condition that all combinations of sites in a molecule have the same distance

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1 - \delta_{ij})\frac{n}{n-1}\rho(G(r) - g(r)). \quad (3.19)$$

Matrix  $\mathbf{\Omega}$  is therefore written as

$$\hat{\mathbf{\Omega}} = \frac{n}{n-1}\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}})(\mathbf{J}_n - \mathbf{I}_n) + \mathbf{I}_n, \quad (3.20)$$

where  $\mathbf{I}_n$  is the identity matrix and  $\mathbf{J}_n$  is the matrix of all ones. The other variables in the RISM-OZ equation are determined straightforwardly. All particle number densities  $\rho_i$  and distribution functions  $g_{ij}$ ,  $h_{ij}$  and  $c_{ij}$  are equal for all  $i$  and  $j$ , which we use to express the matrices in terms of  $\mathbf{I}_n$  and  $\mathbf{J}_n$

$$\rho = \rho\mathbf{I}_n \quad \hat{\mathbf{g}} = \hat{g}\mathbf{J}_n \quad \hat{\mathbf{h}} = \hat{h}\mathbf{J}_n \quad \hat{\mathbf{c}} = \hat{c}\mathbf{J}_n. \quad (3.21)$$

Inserting equations (3.20) and (3.21) in (3.17) we find that the matrix equation reduces to a single equation because of the identity  $\mathbf{J}_n\mathbf{J}_n = n\mathbf{J}_n$

$$\hat{h}\mathbf{J}_n = \left( \frac{n}{n-1}\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}})(n-1) + 1 \right) \hat{c} \left( \left( \frac{n}{n-1}\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}})(n-1) + 1 \right) + n\rho\hat{h} \right) \mathbf{J}_n \quad (3.22)$$

from which follows

$$\hat{c} = \frac{\hat{h}}{(1 + n\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}}))^2 + (1 + n\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}}))n\rho\hat{h}}. \quad (3.23)$$

Note that this result is similar to the PRISM-OZ relation, which is often employed for polymers.[135] We can use this relation with the HNC closure in equation (3.3) for an initial potential guess. For a Newton update in the HNCN/IHNC scheme, we take the derivative with respect to the RDF

$$\frac{d\hat{c}}{d\hat{g}} = \frac{\hat{1}}{(1 + n\rho(\hat{\mathbf{G}} - \hat{\mathbf{g}}) + n\rho\hat{h})^2}. \quad (3.24)$$

Here we ignore that the intramolecular correlation ( $\hat{G} - \hat{g}$ ) is to some degree also a function of  $\hat{g}$ . We assume that intra- and intermolecular degrees of freedom are uncoupled. With this equation we have all we need for a Newton potential update. For  $G - g = 0$ , i.e. no intramolecular correlation, equation (3.24) reduces to equation (3.9).

### 3.2.4 RDF Extrapolation

The sampling of the RDF from a trajectory is, together with the MD simulation, the computationally most demanding task in iterative coarse-graining. It grows significantly with a larger range since the number of particles considered grows roughly with the volume. Here we present a simple scheme for extrapolating an RDF with integral equation theory.

The direct correlation function  $c$  decays much faster to zero than the total correlation function  $h$ . [94] With the OZ equation we can first calculate  $c$  from short-range data for  $h$  using equation (3.6), extrapolate the result with zeros, and then use the OZ equation again to obtain  $h_{\text{ext},0}^\dagger$

$$h_{\text{ext},0} = B^g h_{\text{ext},0}^\dagger = B^g \mathcal{F}^{-1} \left( \frac{\mathcal{F}(B c)}{1 - \rho \mathcal{F}(B c)} \right). \quad (3.25)$$

Here,  $B$  is an operator that appends zeros which numerically equals a unit matrix stacked on top of a block of zeros. The ratio of rows to columns equals the factor by which the range is expanded.  $B^g$  is the generalized inverse of  $B$  which cuts off the tail of a function. The superscript  $\dagger$  denotes that a function is defined on the new longer range. Upon a simple application of equation (3.25) we find that  $h_{\text{ext},0}^\dagger$  does generally look like an extrapolation of  $h$  but has two issues. It deviates from  $h$  on the short range and has a discontinuity at the transition point as depicted in fig. 3.1. The appended tail of  $c$  influences all parts of the total correlation function because of the convolution in the OZ equation.

We now aim to find an improved direct correlation function  $c_{\text{ext}}$  that when plugged in equation (3.25) will result in  $h_{\text{ext}}^\dagger$  which matches  $h$  in the first region. We find that equation (3.25) is not invertible, which makes this an inverse problem. In order to solve it we apply Newton's method where for iteration  $l$  we alternately apply

$$\begin{aligned} c_{\text{ext},l+1} &= c_{\text{ext},l} - \left( \frac{dh_{\text{ext}}}{dc_{\text{ext}}} \right)^{-1} (h_{\text{ext},l} - h) \\ \frac{dh_{\text{ext}}}{dc_{\text{ext}}} &= B^g \mathcal{F}^{-1} \left( \frac{1}{(1 - \rho \mathcal{F}(B c_{\text{ext}}))^2} \right) \mathcal{F} B \end{aligned} \quad (3.26)$$

and equation (3.25) for obtaining  $h_{\text{ext},l+1}$  from  $c_{\text{ext},l+1}$ . We find this to converge within few iterations and giving the expected results, which are exemplarily shown in fig. 3.1.

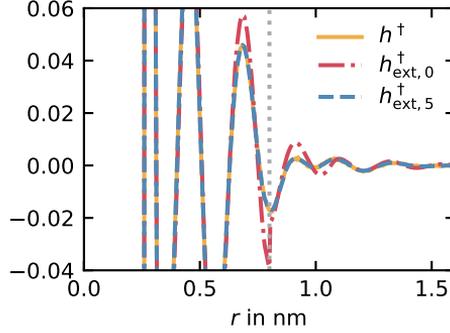


Figure 3.1: The true long-range total correlation function  $h^\dagger$  of the center of mass of SPC/E water and the naive result  $h_{\text{ext},0}^\dagger$  from extending the direct correlation function with zeros. Also shown is the extrapolation result  $h_{\text{ext},5}^\dagger$  from running five Newton iterations to fit  $h$ . The dotted, grey line marks the end of the range of  $h$ , from which  $h_{\text{ext},0}^\dagger$  and  $h_{\text{ext},5}^\dagger$  are calculated.

In order to use this method for the molecular case with  $n$  beads as described before, the following two equations are derived from equation (3.23) for the calculation of  $h_{\text{ext}}$  and its derivative in the Newton scheme

$$h_{\text{ext}} = \mathbf{B}^g \mathcal{F}^{-1} \left( \frac{(1 + n\rho\mathcal{F}(\mathbf{B}(G - g)))^2 \mathcal{F}(\mathbf{B}c)}{1 - n\rho(1 + n\rho\mathcal{F}(\mathbf{B}(G - g)))\mathcal{F}(\mathbf{B}c)} \right) \quad (3.27)$$

$$\frac{dh_{\text{ext}}}{dc_{\text{ext}}} = \mathbf{B}^g \mathcal{F}^{-1} \left( \frac{(1 + n\rho\mathcal{F}(\mathbf{B}(G - g)))^2}{(1 - n\rho(1 + n\rho\mathcal{F}(\mathbf{B}(G - g)))\mathcal{F}(\mathbf{B}c_{\text{ext}}))^2} \right) \mathcal{F}\mathbf{B}. \quad (3.28)$$

## 3.3 Method

### 3.3.1 HNCN and HNCNGN in VOTCA

The two new coarse-graining methods HNCN and HNCNGN described above have been implemented in the Versatile Object-oriented Toolkit for Coarse-graining Applications (VOTCA) [87, 136] software. For HNCN, all variants discussed in section 3.2.1 are available and can be specified in the VOTCA input file. The usage is similar to IBI and IMC, Python and NumPy are used internally for the vector operations. The code allows for the generation of a potential guess based on integral equations (equations (3.3) and (3.7)). A potential guess for molecular systems of identical beads can be generated using equation (3.23). Potential updates are calculated with the HNCN, IHNC, and HNCNGN method (equations (3.10), (3.11), and (3.12)). We also modify VOTCA's `csg_stat`

---

---

to compute the correlation function  $G(r)$ , which is needed for the molecular case. An alternative to the HNC closure, the Percus-Yevick closure

$$g = \frac{c}{1 - e^{\beta u}} \quad (3.29)$$

can be used to derive related update schemes, which are also implemented but not tested thoroughly. Some preliminary tests showed very similar results to the HNC closure for the systems in this paper. The pressure constraint for HNCGN is implemented as an elimination with the algorithm as described by Gander et al.[78] In the current form, the code can be straightforwardly extended to other constraints if they can be expressed in terms of the force  $F(r)$  and the RDF  $g(r)$ .

There is also a new *power* extrapolation scheme for the potential in the core region where the RDF is zero. It fits a power function  $U_{\text{fit}} = ar^b$  to the Boltzmann inverse of  $g_{\text{tgt}}$  from the first  $r$  where  $g_{\text{tgt}}$  is larger than a threshold (default:  $1 \times 10^{-10}$ ) to its first maximum. The potential is then extrapolated in the region  $r = 0$  to the first  $r$  where the potential is convex or  $g_{\text{tgt}}$  is above  $1 \times 10^{-2}$ .

The code resides currently in a fork of VOTCA at [https://gitlab.com/cpc\\_group/csg](https://gitlab.com/cpc_group/csg) but we aim to get it into the main repository.

### 3.3.2 Water, Hexane, Naphthalene

We have tested our methods on simple LJ systems and found similar results as Delbary et al.[75], so we turn to more relevant molecular systems in this work. To test the new coarse-graining schemes we create reference structures by performing all-atom (AA) simulations of water, hexane, and naphthalene. We run NVT simulations of SPC/E water and OPLS/AA models of hexane and naphthalene.[16, 137, 138] Long-range electrostatics are accounted for with the particle-mesh-Ewald (PME) method. GROMACS 2019 was used for all simulations.[139] For all systems, a box with the correct volume is filled randomly with molecules. It is then energy-minimized with the steepest descent algorithm for  $10^4$  steps. After an equilibration of  $10^5$  steps a production run of  $2 \times 10^6$  steps is performed where every 100th step is written to a trajectory file. A Langevin thermostat with a friction constant of  $2 \text{ ps}^{-1}$  keeps the temperature constant. Further simulation details and the average pressure from the production run are given in table 3.1.

The VOTCA package is used for mapping the AA trajectory and calculating the target distribution functions for the coarse-graining. The mapping schemes used are shown in fig. 3.2. In naphthalene, each bead corresponds to the center of mass of two carbon atoms and one of the central carbon atoms with half weight. A snapshot from the equilibrated AA simulation is taken and mapped to be the starting configuration. The starting non-bonded potentials are generated by equation (3.7), with the target distribution over a range of twice the cut-off. This ensures that all methods start from the same potential and allows us to compare the potential updates in the first iteration in detail. For the bonded

Table 3.1: Parameters used in the AA NVT simulations and the resulting average pressure.

	water	hexane	naphthalene
Nr. of molecules	5000	2000	2000
Temperature in K	298.15	273.15	373.15
Density in $\text{g cm}^{-3}$	0.998	0.67	0.99
Timestep in fs	2	1	1
Cut-off in nm	0.8	1.2	1.2
NVT pressure in bar	259	77	311

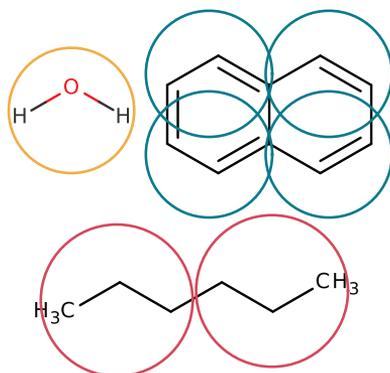


Figure 3.2: Mapping schemes for water, hexane, and naphthalene. Bead positions are determined by the center of mass of atoms within it. The central carbon atoms in naphthalene belong to two beads at the same time.

potentials (one bond in hexane, two bond-types, one angle-type, and two dihedral-types in naphthalene) the target distribution is Boltzmann inverted to obtain a starting potential. For each iteration the system is run with a timestep of 4 fs and a cut-off of 0.8 nm for water and 1.2 nm for hexane and naphthalene. MD simulations are run for  $7 \times 10^4$  ( $3.1 \times 10^5$  for IMC) steps where every 10 steps the positions are saved in a trajectory, of which the first  $1 \times 10^4$  are discarded. This means, that five times the amount of frames goes into sampling the IMC matrix, compared to the distributions for the other methods. For HNCN, HNCN (jc), and HNCGN the RDF for the potential update at each iteration is determined up to twice the cut-off. In the core region where the potential update is undefined due to the RDF being zero the potential is extrapolated with an exponential function. The simulation settings are otherwise equal to the AA simulations.

While we vary the methods for the non-bonded potential update, we consistently use

---

---

IBI for the internal degrees of freedom. This ensures that bond and angle distributions are precisely reproduced which simple Boltzmann inversion cannot guarantee. We do however have to scale the potential updates before adding them to the potential for the previous iteration. This is done for naphthalene, where we use 0.5 for angles and 0.25 for dihedrals. The scaling is justified by the interdependence of multiple bonded potentials due to the small ring. In ring-free molecules, bonded potentials are normally relatively independent. However, in a four-ring a small change in all four angle potentials will make the whole ring much stiffer. Therefore, to make the angle distribution narrower the potential needs to be changed only a fraction of the IBI update. If applied without the scaling, we observe divergence in the bonded potentials and the simulation crashes after some iterations.

### 3.3.3 Methanol in Water

Methanol-water simulations are prepared with seven different mole fractions of methanol: 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9. The AA simulations employ the OPLS/AA model for methanol and the TIP4P model for water.[13, 137] A total of 8000 molecules with eight different mole fractions are equilibrated and simulated in an NPT ensemble. Molecules are inserted randomly into a box with density  $1 \text{ g}^3 \text{ cm}^{-1}$ . They are successively energy minimized for  $5 \times 10^3$  steps, run with a velocity rescaling thermostat with a stochastic term for  $1 \times 10^4$  steps, and run with an additional Berendsen barostat for  $1 \times 10^5$  steps. The timestep is 1 fs and the temperature 298.15 K. The cut-off for the LJ interactions is 1.2 nm and PME is used for the long-range electrostatics. The production run is performed with a Nosé-Hoover thermostat ( $\tau_T = 2$  ps) and Parinello-Rahman barostat ( $\tau_p = 10$  ps). Since we are interested in the methanol-methanol distribution function, we simulate longer for dilute systems. The production run has therefore  $\frac{2 \times 10^5}{x(\text{methanol})}$  steps. Positions of the atoms in methanol are written every ten steps after the first 0.5 ns. An RDF is calculated from the center of mass of each methanol molecule using VOTCA.

For each composition, we also determine the osmotic pressure from atomistic simulation. We therefore closely follow the OPAS method.[140] It is based on a simulation of a box elongated in z-direction where the solute is kept in a defined region by two semipermeable membranes normal to the z-axis. From a pre-run with a total pressure of 1000 bar the osmotic pressure is first estimated, such that in the production run the pressure outside the mixed slab can be set to 1 bar. We depart slightly from the original method in that we employ a Parinello-Rahman barostat that scales the box in x- and y-direction during the pre- and the production-run. Thereby the box is scaled perpendicular to any forces exerted by the walls and interference is avoided. Equilibration is performed in the same procedure as described above for the bulk simulations. We test two force constants  $k_w$  for the half-harmonic potentials that form the semipermeable membranes,  $500 \text{ kJ mol}^{-1} \text{ nm}^{-2}$  and  $4000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ . Methanol molecules are attached to a virtual site at their center of mass. The membrane force is determined by the position of the

---

---

virtual site and the force is distributed on the atoms of the molecule. Inside the mixture slab, some of the methanol molecules adsorb to the semipermeable walls and also push a bit outside of the confined region. The resulting mole fraction and concentration for each osmotic pressure is determined from the unperturbed region in the middle of the slab, by counting molecules in that region.

We develop an implicit solvent model of methanol in water for each mole fraction. Methanol is mapped to a single bead. We use the HNCGN method with and without a pressure constraint to obtain the CG interaction potentials. The configuration is generated from a mapped configuration of a snapshot of an  $x(\text{methanol}) = 0.9$  AA system with 7200 methanol molecules, which is scaled to match the methanol density of the NPT run. This results in large boxes for the low mole fractions. Therefore we need again more sampling for the “dilute” systems to obtain meaningful RDFs. We run each coarse-graining iteration for  $1 \times 10^4$  equilibration steps and  $\frac{4 \times 10^4}{x(\text{methanol})}$  production steps of which every tenth is saved.

## 3.4 Results and Discussion

### 3.4.1 Potential Guess

We start by examining the potential obtained from equation (3.7) and (3.23) which we use for all performed iterative coarse-graining runs to start from. In fig. 3.3 we compare it to the Boltzmann inverted (BI) target distribution, i.e. the potential of mean force, which is the common choice in coarse-graining for the starting potential. For the HNC potential of hexane and naphthalene, the intramolecular distribution obtained from the mapped AA simulation is used. We find that the HNC potential is always more repulsive for all three molecules compared to the BI potential. It does result in a very good RDF when compared with the BI potential, especially for water, where the BI potential generates too much structuring. For naphthalene both potentials, even though very different in shape, produce almost the same RDF. This confirms the common observation that while there might be only one optimal potential, the RDF is very insensitive to certain changes in the potential.

### 3.4.2 Jacobian Comparison

To analyze the quality of the update schemes, we compare the Jacobians of each Newton method. In fig. 3.4 the Jacobians from the first iteration of water with the respective methods are shown. Each element represents the derivative  $\frac{\partial g_\alpha}{\partial u_\gamma}$  for the change of the RDF at  $\alpha$  by changes in the potential at  $\gamma$ . In the first iteration, the Jacobian contains all information about the update, since for all methods the distance  $g_k - g_{\text{tgt}}$  is the same.

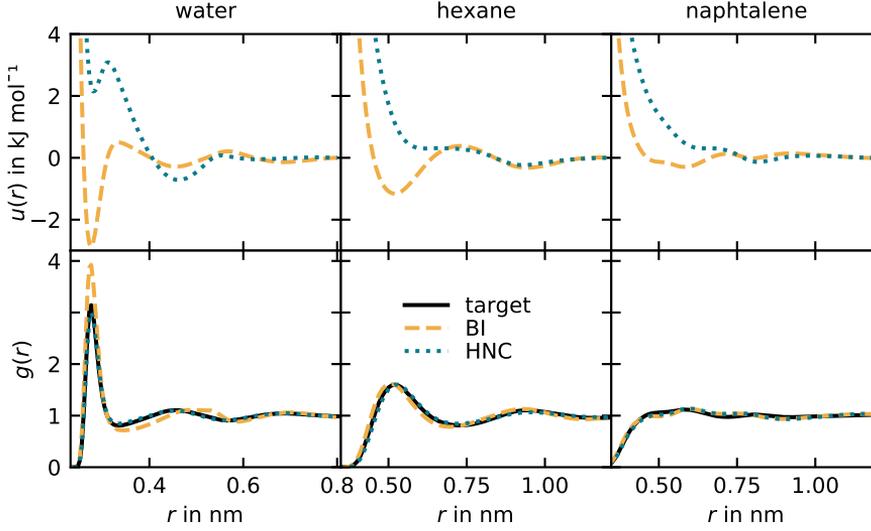


Figure 3.3: Potentials generated from Boltzmann inversion (BI) and HNC inversion of a given target RDF and the resulting distributions from MD simulations.

For IBI it is a diagonal matrix with values

$$\text{diag}(\mathbf{J}_{\text{IBI}}) = \beta((g_k - g_{\text{tgt}}) / \log(g_{\text{tgt}}/g_k)) \quad (3.30)$$

on the diagonal and zero on the off-diagonal elements. The diagonality reflects that IBI can only update the potential based on local information about structure mismatch. For IMC usually the IMC matrix  $\mathbf{A}_{\text{IMC}}$  is written down in terms of a function  $S(r) = 4\pi r^2 N^2 / (2V) g(r)$  where  $N$  is the number of particles and  $V$  is the volume. The elements of the IMC matrix are calculated by

$$\mathbf{A}_{\text{IMC}\alpha\gamma} = \frac{\partial S_\alpha}{\partial u_\gamma} = \beta (\langle S_\alpha \rangle \langle S_\gamma \rangle - \langle S_\alpha S_\gamma \rangle). \quad (3.31)$$

For comparison purpose, the Jacobian with respect to  $g$  can be retrieved by

$$\mathbf{J}_{\text{IMC}\alpha\gamma} = \frac{\partial g_\alpha}{\partial u_\gamma} = (4\pi r_\alpha^2 N^2 / (2V))^{-2} \mathbf{A}_{\text{IMC}\alpha\gamma} \quad (3.32)$$

The Jacobian for the HNCN method is the term in square brackets in equation (3.10), inverted.

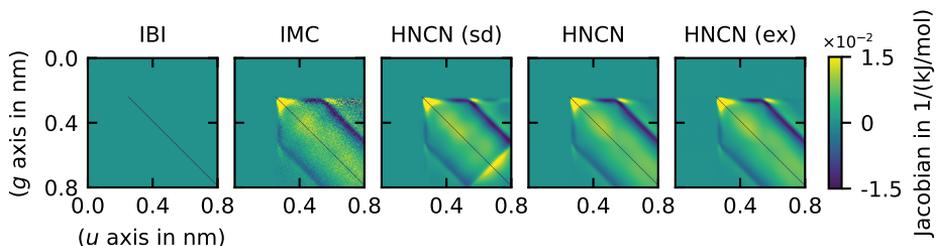


Figure 3.4: Jacobian  $\frac{dg}{du}$  for the first iteration of the SPCE-water coarse-graining from four different methods. For HNCN and HNCN (ex) the Jacobian extends till 1.6 nm in both directions and only the top left quarter is presented.

The diagonal elements of all Jacobians are negative since a positive perturbation of the potential at  $\alpha$  will result in a negative impact in the RDF at the same distance. Except for IBI, other features of the three Jacobians are very similar too, for example the positive region close to the diagonal at about the first peak of the RDF. Physically, this can be thought of as an effect of the potential well. A positive potential perturbation in that region leads to a higher sampling of neighboring points because the remaining potential well is favored. Another motif that is present in both methods is a pair of diagonal stripes of negative values, which are located  $\pm 0.27$  nm from the diagonal. That distance is equivalent to the position of the first peak  $r_{fp}$  in the RDF. It represents that a perturbation of the potential at any distance will make it less likely to find a second particle at that distance which will indirectly make the presence of a third particle at that distance  $\pm r_{fp}$  less likely. The IMC Jacobian is not perfectly smooth, where the noise depends on the number of frames used for sampling the IMC matrix. It is symmetric by construction, see equation (3.31) and fading out to the bottom due to the normalization with  $\frac{1}{r^2}$  in equation (3.32). The Jacobians  $\mathbf{J}_{\text{HNCN (sd)}}$ ,  $\mathbf{J}_{\text{HNCN}}$ ,  $\mathbf{J}_{\text{HNCN (ex)}}$  are much smoother but seem to retain the same structure and symmetry.

The similarity in structure indicates that through integral equations it is well possible to construct a Jacobian just from RDFs.  $\mathbf{J}_{\text{HNCN (sd)}}$  shows an interesting artifact in the bottom right corner: a stripe of positive values orthogonal to the matrix diagonal. The same motif is present in  $\mathbf{J}_{\text{HNCN}}$  and  $\mathbf{J}_{\text{HNCN (ex)}}$ , but outside the shown sector of the matrix. On closer inspection, there are more, weaker but similar motifs that go perpendicular to the diagonal. We account those artifacts to the numerical deconvolution of  $c$  and  $h$ . In the OZ equation (3.2) the total correlation function  $h$  is input and output of the convolution. One aspect of the convolution of two distributions is that the output will have a larger range than the inputs. The situation with the OZ equation is best illustrated

by its recursive expansion

$$h = c + \rho h * c = c + \rho c * c + \rho^2 c * c * c + \dots \quad (3.33)$$

So even if the direct correlation function decays very fast, which it usually does, it will generate non-negligible values for the total correlation function on a longer range than  $c$  has. Therefore, by calculating  $c$  from  $h$  with a short-ranged  $h$  one will produce an erroneous estimate for  $c$ . The same argument can be made for the calculation of  $\frac{dc}{dg}$  from  $h$ . We think this effect is also the reason why Heinen calculates the structure factor in a way that uses effectively the whole simulation box.[127] The original HNCN method avoids this by calculating a longer RDF and cutting away the possibly erroneous part of the potential. The HNCN (jc) method as introduced in section 3.2.1 also avoids the artifact by cutting the Jacobian, before multiplication with the RDF distance. The Jacobian of the HNCN and the HNCN (jc) method are equal, so we only show the former in fig. 3.4. Finally, the variant where the RDF has been extrapolated, HNCN (ex), shows a Jacobian indistinguishable from the HNCN (jc) method.

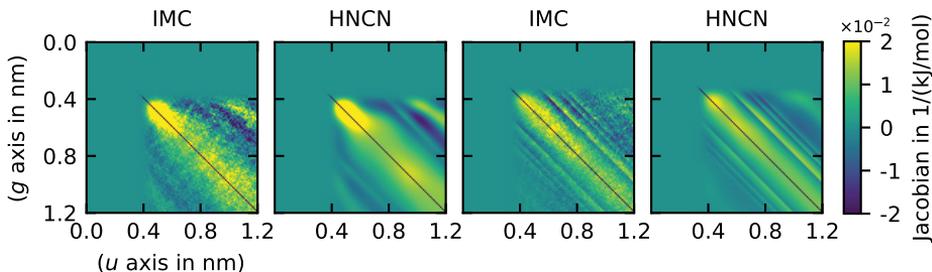


Figure 3.5: Jacobians for the first iteration of the hexane (left two) and naphthalene (right two) coarse-graining with IMC or the HNCN method. For HNCN the Jacobians extend till 2.4 nm in both directions and only the top left quarters are presented.

To confirm the newly derived update scheme for symmetric molecules we also compare the Jacobians of hexane and naphthalene with IMC in fig. 3.5. Due to the absence of noise, finer details are recognizable in the HNCN Jacobian, such as sharp lines, that originate from the peak like intramolecular distribution function. The HNCN Jacobian is indeed very similar to the IMC result. This confirms, that a Newton update based on the RISM equation does sufficiently approximate the exact Newton update. It also implies that the approximation made in equation (3.19) for naphthalene, namely the equivalence of off-diagonal elements in  $G_{ij}$ , is reasonable.

We cannot present a Jacobian calculated from HNCN (ex) at this point. The reason is that the Newton scheme for RDF extrapolation, as described by equations (3.27) and

(3.28), fails to converge for the distribution generated by the potential guess for both hexane and naphthalene. Interestingly, it converges for the target distribution in both cases. We believe this is related to the inadequateness of the assumption of a fast decay of the direct correlation function for the molecular case. The RISM-OZ equation (3.17), in contrast to the simple OZ equation (3.2), is known to not be fully consistent with a short-ranged direct correlation function.[94, 141]

### 3.4.3 Convergence

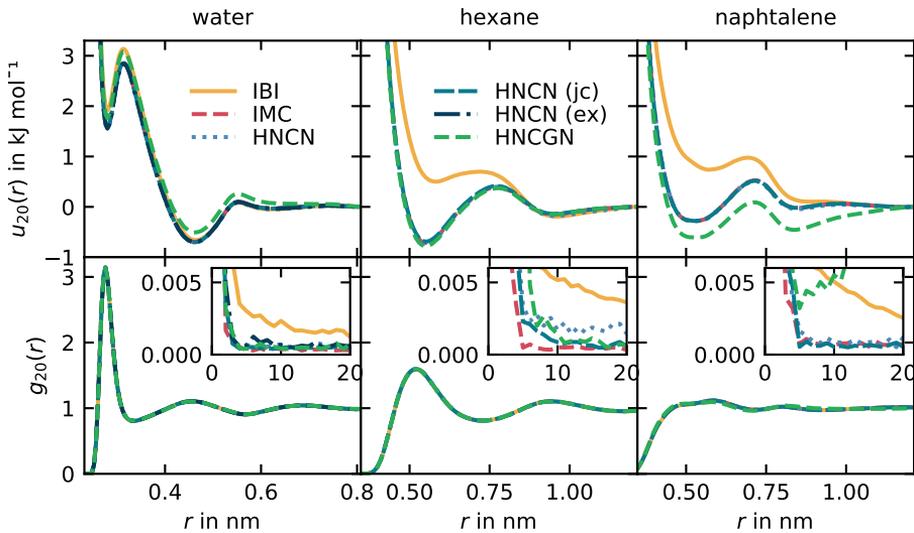


Figure 3.6: Potentials and distributions up to the cut-off after 20 iterations of four different coarse-graining methods. The insets in the bottom row show the convergence of the data fit  $\delta$  versus the iteration number. The target RDF is also plotted, but is not visible because it is hidden by the graphs of the other distributions.

We now compare the output of six different coarse-graining methods: (i) IBI, (ii) IMC, (iii) HNCN, (iv) HNCN (jc), (v) HNCN (ex), and (vi) HNCGN. Bonded potentials have in any case been updated using IBI and have converged within few iterations in any case, so we are not showing them here. In fig. 3.6 we show the potentials and distributions

---

---

obtained after 20 iterations. The convergence of the RDF, given by the data fit measure

$$\delta = \sqrt{\frac{1}{r_{\text{co}}} \int_0^{r_{\text{co}}} (g_k - g_{\text{tgt}})^2 dr}, \quad (3.34)$$

is also shown.[87] The upper integration limit is the cut-off  $r_{\text{co}}$ . HNCN (ex) results are only present for water since, as mentioned in the last section, the RDF extrapolation fails for hexane and naphthalene. Results for the HNCN (sd) method are missing for all three molecules because the iterative method would produce erroneous potentials: those would be strongly oscillating over their whole range or have very deep minima which would at some point crash the MD simulation. Since the other HNCN methods converge, we contribute the convergence failure to the artifacts in the Jacobian as discussed previously. This limits the applicability of the naive HNCN (sd) method. If short cut-offs are wanted, and this is typically the case in coarse-graining, an Ansatz with longer RDF for the Jacobian has to be used to obtain meaningful results.

All methods produce the same distribution within line thickness after 20 iterations, the potentials however differ. The IBI potentials are much more repulsive in hexane and naphthalene, but we did not test if they would eventually converge to the IMC potential. The IMC, HNCN, and HNCN (jc) results are almost identical, and also converge to a similar  $\delta$  value. The convergence speed is fastest with IMC only requiring 3-8 iterations till  $\delta$  flattens out. HNCN, HNCN (jc), and HNCN (ex) converge within 4-11 iterations, a bit slower than IMC but much faster than *IBI*. This includes hexane and naphthalene which shows that RISM based coarse-graining works for the examined molecular systems.

We find that for hexane HNCN achieves less accuracy than HNCN (jc), i.e. the converged  $\delta$  value is higher. HNCN (jc) only uses a cut-out of the Jacobian, while HNCN uses the inverse of the full matrix including the artifact. Even though the tail of the potential update is cut off, it seems the artifact still influences the short-ranged potential update.

One can compare the performance by looking at the amount of time used for each iteration. For water we find 1.5 min for IBI, HNCN (sd), and HNCN (ex), 6 min for IMC, and 7 min for HNCN, HNCN (jc), and HNCN (ex), on a 24 core AMD Opteron 6174. The bottlenecks are the MD simulation, the RDF calculation, and the IMC matrix calculation. HNCN and HNCN (jc) were performed with the RDF calculated on double the range than HNCN (sd), which has a large performance impact. HNCN (ex) has the same input as HNCN (sd) and extrapolates the RDFs to the doubled range on the fly. This results in HNCN (ex) being the fastest method to converge in total computational time, beating IMC by a factor of four.

When looking in more detail at the convergence behavior we find that the RDF oscillates slightly around the target distribution. Oscillations around the solution are expected from a Newton method with a slight error in the derivative. For IMC this error in the Jacobian is statistical. For HNCN (jc) it is statistical, caused by noise in the

distributions, and systematic due to the HNC closure not being exact. It follows that if one is interested in improving the RDF to a very precise degree, the way to go is using IMC updates with sufficient statistics in the IMC matrix. For practical purposes, the precision of HNCN (jc) should be more than satisfying.

The HNCNGN results are less uniform. We find that it converges similarly fast as HNCN for water and hexane, but to slightly different potentials, once more repulsive, once more attractive. For naphthalene, the situation is worse and the potential oscillates largely and never converges. To understand if artifacts in the Jacobian can be the source of this, we have to know which parts of it are used. HNCNGN by default can use the RDF on a longer range than the potential update is made on through matrix  $\mathbf{A}_0$  in equation (3.12). Due to the block of zeros in the lower part of  $\mathbf{A}_0$  it cuts out parts from the right of the Jacobian when multiplied with it. This reflects the idea of the Gauss-Newton Ansatz, where a short-ranged potential is updated based on a longer RDF and non-square shape allows for the addition of constraints. We depict this in fig. 3.7. The artifacts, which

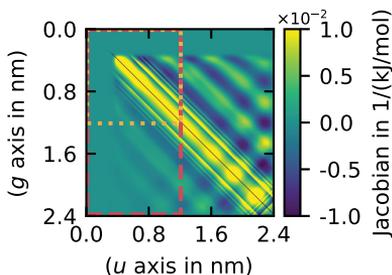


Figure 3.7: The Jacobian obtained from integral equations for the first coarse-graining iteration of naphthalene. The orange dashed rectangle shows the region that is used for the HNCNGN update, whereas the yellow dotted square shows the cut out that HNCN (jc) is using.

are visible as “waves” perpendicular to the diagonal and emerging from the lower right corner, are present in the HNCNGN region. We cannot with full certainty link the issues in convergence with HNCNGN in naphthalene to these artifacts in the Jacobian, but since they have shown to be the root cause with the HNCN (sd) method it seems likely. The main artifact lies outside of the region used, which probably explains why HNCNGN fails to converge only sometimes. A concluding test would be to calculate the Jacobian on an even larger range, such that the part that is used by HNCNGN is uninfluenced by the artifacts, but this we have not implemented. The difference in the potentials obtained for water and hexane can be explained by that HNCN and HNCNGN are optimizing different residues. HNCNGN finds the best potential for the RDF on a range twice as long as HNCN does. So the difference in the potential is to be expected since more and different target information goes into the HNCNGN method.

### 3.4.4 Pressure Matching

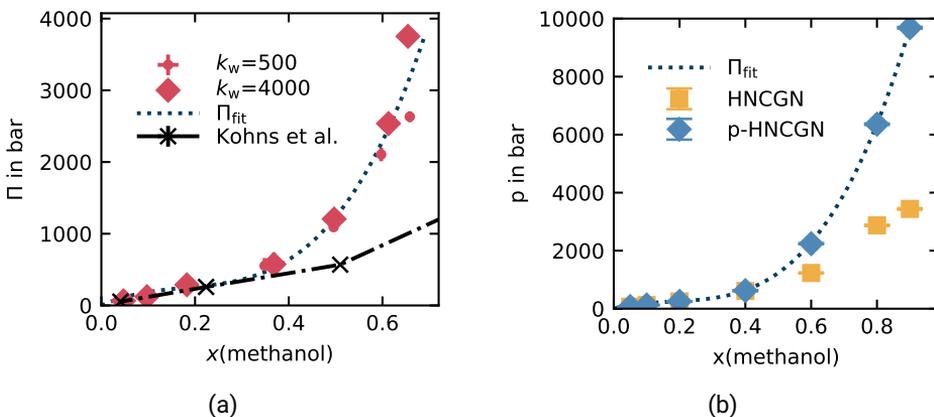


Figure 3.8: (a) Osmotic pressures obtained by the OPAS method from all-atom slab simulations with two different wall force constants  $k_w$  in  $\text{kJ mol}^{-1} \text{nm}^{-2}$ . The polynomial fit function (eq. (3.35)) is included as dotted line. The literature data from Kohns et al. are shown as crosses and connected by dashed-dotted lines to guide the eye. (b) The pressure values resulting from NVT simulations of CG models of methanol in implicit water.

We now present the results for the methanol-water mixtures. Figure 3.8a displays the osmotic pressures  $\Pi$  determined by the OPAS method. The osmotic pressure is found to increase strongly up to several thousand bars when the fraction of methanol goes up. The two different wall force constants seem to have little systematic influence on the osmotic pressure, but our error bars calculated from block averaging over time are probably too low. Literature values from Kohns et al. are significantly lower for higher mole fractions.[142] However, they use the TIP4P/2005 water model and a united-atom methanol model whereas we use the TIP4P water model and the OPLS-AA methanol model, so differences have to be expected. We take the data for  $\Pi$  from both force constants and fit them with a polynomial of third order without zero order term,

$$\Pi_{\text{fit}}(x) = ax^3 + bx^2 + cx. \quad (3.35)$$

The resulting fit parameters are  $a = 24\,070$  bar,  $b = -12\,960$  bar, and  $c = 2930$  bar; this fit is also shown in fig. 3.8a.

To illustrate the ability of our scheme of matching these pressures we compare (i) HNCGN without constraints and (ii) HNCGN with pressure constraint (p-HNCGN) on the potential updates. We use the beforementioned polynomial  $\Pi_{\text{fit}}$  to set the target pressure

for the p-HNCGN method at the given mole fractions. Osmotic pressures for  $x > 0.65$  are extrapolated values because of the lack of corresponding OPAS data for higher mole fractions.

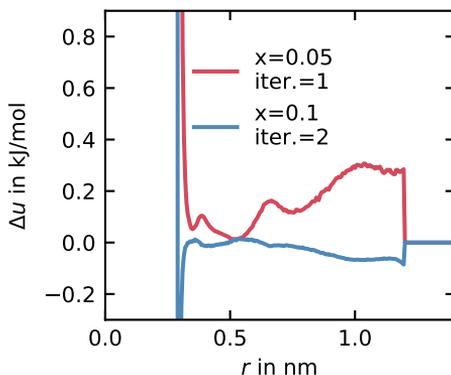


Figure 3.9: Two exemplary potential updates that show spikes near the core region and jumps before the cut-off arising from using applying the p-HNCGN method without post-processing.

Upon straightforward implementation of the p-HNCGN method, we embrace two problems that can cause unphysical potentials. The first is a jump in the potential update right before the cut-off as seen in the potential update of iteration one for  $x = 0.05$  in fig. 3.9. This is related to the pressure constraint since the sudden jump in the potential creates a huge force at that distance, which increases the pressure while having a small effect on the structure. This effect is stronger in the more dilute systems. To prevent this, we implement a simple fix, where the whole potential is shifted, such that it is continuous at the cut-off. However, we do not use it here, since we find the final potentials only have this jump for  $x = 0.05$  and it was relatively small. Secondly, there is a negative spike that appears when the pressure is corrected down. It appears at the end of the core region, where the current and target RDF are close to zero. We show an example in fig. 3.9. This can create a very narrow and deep potential well, which crashes the MD simulation. We circumvent this by extrapolating the potential, e.g. by the *power* scheme described in section 3.3.1. This scheme, contrary to the standard scheme, substitutes a small part of the repulsive region of the calculated potential. With this tweak, we find the coarse-graining iterations to be stable and to converge after approximately five iterations.

The potentials and distributions obtained from the p-HNCGN method after 20 iterations are shown in fig. 3.10. The potentials differ and get more repulsive with a growing mole fraction of methanol. For implicit solvent systems, one cannot expect there to be a common potential for all concentrations, since the influence of the water is incorporated

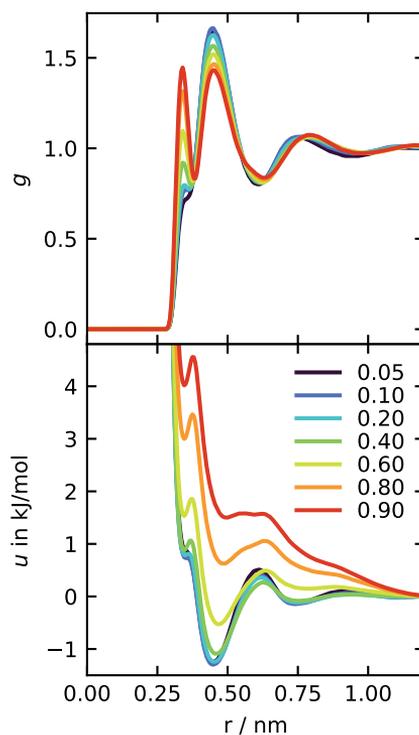


Figure 3.10: RDFs and potentials for CG methanol obtained with the p-HNCGN method.

in the potential. However, we find that for concentrations lower than  $x \leq 0.4$  there are only small differences in the potentials. This indicates that below this concentration methanol molecules interact in a similar way, independent from the concentration. Above  $x = 0.4$  we find the potentials to become much more repulsive and increasingly showing two distinct minima, one at 0.35 nm and one at 0.45 nm. The pressures obtained from the HNCGN and p-HNCGN potential are shown in fig. 3.8b and in table 3.2. The error estimates are smaller for the more dilute systems because they have been simulated longer. From the graphic representation, we can see that the target pressure is met very well, where the unconstrained HNCGN method results in too low pressure.

Table 3.2: Pressures in bar of CG models of methanol water mixtures from two different methods.

$x$ (methanol)	target	HNCGN	p-HNCGN
0.05	117	67(1)	75(1)
0.1	187	125(3)	129(3)
0.2	260	251(6)	253(6)
0.4	639	611(13)	623(12)
0.6	2291	1232(18)	2240(18)
0.8	6373	2872(22)	6360(23)
0.9	9686	3439(26)	9685(26)

### 3.5 Conclusion and Outlook

We have demonstrated the applicability of the Newton methods based on integral equation theory for molecular coarse-graining. The results show that from the OZ equation the Jacobian for a Newton update can be deduced which is equivalent to the IMC Jacobian. The new HNCN methods have properties similar to IMC: rapid convergence in under 10 iterations and the same resulting potential, but it does not require the sampling of the IMC matrix. For short potential cut-offs, which are prevailing in molecular coarse-graining, we find the HNCN Jacobian to have artifacts if the RDF is not sampled on a longer range. They are explained by the numerical deconvolution of the OZ equation. Those artifacts are found to prohibit convergence of the method and distributions have to be evaluated beyond up to around the double cut-off to construct a valid Jacobian. Using HNCN with a longer RDFs makes it as slow as IMC, but when we use a physically motivated scheme to extrapolate the RDFs we obtain four times faster convergence than IMC for water.

An extension to multiple-site representations of molecular liquids based on the RISM-OZ equation has been proposed and demonstrated to work well for CG models of liquid hexane and naphthalene. Again, the resulting potentials are similar to IMC results and convergence speed and accuracy is comparable. This is a first step towards applying iterative integral equation coarse-graining methods in general systems. The same approach could also be used for coarse-graining of polymer melts if the intramolecular distribution function is approximated to be equal between all bead combinations. Existing uses of the OZ equation in polymer coarse-graining might profit from a Newton formulation for computing improved pair potentials.

The HNCGN method has been used to derive models for a single bead CG methanol model in implicit water at different concentrations. The p-HNCGN method can match the osmotic pressure that was previously determined by OPAS simulations. The implementation of pressure constraints by elimination from the Gauss-Newton problem proves

to be a powerful tool to enforce additional conditions on the potentials with only minor numerical pitfalls. The formulation could in this form also be used with the IMC method and would probably be more consistent than the common ramp corrections.

We note that IBI has been extended to inhomogeneous systems which improves the structure of the phase boundary. [143] This extension is non-trivial for the methods studied in this work as homogeneity is a requirement for integral equation theory.

## Acknowledgments

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 233630050 - TRR 146 on “Multiscale Simulation Methods for Soft Matter Systems”. The authors thank Viktor Klippenstein and Swaminath Bharadwaj (both Technische Universität Darmstadt) for discussions on aspects of this work.

## Data Availability

The computer code that supports the findings of this study is openly available at <https://dx.doi.org/10.5281/zenodo.4290469>.

## Appendix: Intramolecular Distribution Function

Here we show how in a molecule of  $n$  identical beads the  $G_{ij} - g_{ij}$  relates to  $G - g$ . We first write down the summation over  $N$  molecules with indices  $g$  and  $h$  for the explicit case

$$\rho g_{ij}(r) = \left\langle \frac{1}{N} \sum_g \sum_h^N (1 - \delta_{gh}) \delta(r - |\mathbf{r}_{gi} - \mathbf{r}_{hj}|) \right\rangle \quad (3.36)$$

$$\rho G_{ij}(r) = \left\langle \frac{1}{N} \sum_g \sum_h^N (1 - \delta_{gh} \delta_{ij}) \delta(r - |\mathbf{r}_{gi} - \mathbf{r}_{hj}|) \right\rangle. \quad (3.37)$$

Here,  $\mathbf{r}_{gi}$  is the position of site  $i$  on molecule  $g$ . Therefore we obtain for the difference of the two distribution functions

$$\begin{aligned} \rho(G_{ij}(r) - g_{ij}(r)) &= \left\langle \frac{1}{N} \sum_g \sum_h^N -\delta_{gh} (\delta_{ij} - 1) \delta(r - |\mathbf{r}_{gi} - \mathbf{r}_{hj}|) \right\rangle \\ &= (1 - \delta_{ij}) \langle \delta(r - |\mathbf{r}_i - \mathbf{r}_j|) \rangle_{\text{mol}}. \end{aligned} \quad (3.38)$$

$\langle \rangle_{\text{mol}}$  indicates an average over all molecules. This reflects that the intramolecular distribution function is independent of intermolecular correlations. If all combinations  $i$  and  $j$  have the same distance distribution  $e(r)$  we can simplify to

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1 - \delta_{ij})e(r) \quad (3.39)$$

Now we assume molecules with  $n$  equal sites. The density of those sites is  $n\rho$ . To obtain the site independent expression we need to sum over all sites

$$\begin{aligned} n\rho(G(r) - g(r)) &= \left\langle \frac{1}{nN} \sum_g \sum_h \sum_i^n \sum_j^n -\delta_{gh}(\delta_{ij} - 1)\delta(r - |\mathbf{r}_{gi} - \mathbf{r}_{hj}|) \right\rangle \\ &= \left\langle \frac{1}{n} \sum_i^n \sum_j^n (1 - \delta_{ij})\delta(r - |\mathbf{r}_i - \mathbf{r}_j|) \right\rangle_{\text{mol}}. \end{aligned} \quad (3.40)$$

Again we assume all distance distributions to be equal  $e(r)$  and obtain

$$n\rho(G(r) - g(r)) = (n - 1)e(r). \quad (3.41)$$

By comparing equations (3.39) and (3.41) we find

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1 - \delta_{ij})\frac{n}{n-1}\rho(G(r) - g(r)) \quad (3.42)$$

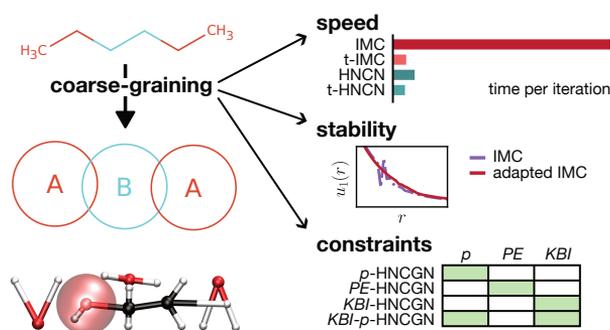
Note that this is an approximation, if not all sites have the same distance distribution. For example in a rectangular molecule  $G_{12}$  would be different from  $G_{13}$ . Equation (3.42) is therefore only exact for molecules where all sites have the same internal distances, i.e. a dumbbell, equilateral triangle, and tetrahedron.



---

## 4 Stability, Speed, and Constraints for Structural Coarse-Graining in VOTCA

---



### Abstract

Structural coarse-graining involves the inverse problem of deriving pair potentials that reproduce target radial distribution functions. Despite its clear mathematical formulation, there are open questions about the existing methods concerning speed, stability, and physical representability of the resulting potentials. In this work, we make progress on several aspects of iterative methods used to solve the inverse problem. Based on integral equation theory, we derive fast Gauss-Newton schemes applicable to very general systems, including molecules with bonds and mixtures. Our methods are similar to inverse Monte Carlo in terms of convergence speed and have a similar cost per iteration as iterative Boltzmann inversion. We investigate stability problems in our schemes and in the inverse Monte Carlo method and propose modifications to fix them. Furthermore, we establish how the pair potential can be constrained at each iteration to reproduce the pressure, Kirkwood-Buff integral, or the enthalpy of vaporization. We demonstrate the potential of our approach in deriving coarse-grained force fields for nine different solvents and their mixtures. All methods described are implemented in the free and open VOTCA software framework for systematic coarse-graining.

---

---

## 4.1 Introduction

Coarse-grained molecular simulations present the challenge of representing a molecular system with a large number of atomistic degrees of freedom by a simple model with a significantly reduced number of degrees of freedom. Its main applications lie in polymer and biomolecular science where the time- and length scale are out of reach for atomistic molecular dynamics (MD) simulations.[27, 40, 144, 145] An important class of coarse-grained (CG) models is particle-based, with pair potentials realizing repulsion and attraction between groups of atoms combined into so-called beads. Several methods exist to obtain CG pair potentials: Top-down methods parametrize the CG potentials based on experimental data.[42] Bottom-up methods derive potentials from some reference atomistic model with a preexisting force field or ab initio MD. Examples for the latter are relative entropy (RE), force-matching, and machine learning.[56, 80, 146] Two methods, inverse Monte Carlo (IMC) and iterative Boltzmann inversion (IBI), explicitly optimize pair potentials to match the radial distribution function (RDF) of the mapped atomic reference.[50, 52] The RE method has been proven to be equivalent to IMC if pair potentials are optimized by a Newton scheme.[91, 92]

The mathematical classification of the structural coarse-graining problem is that of an inverse problem. Only a complicated forward function from the potential to the RDF exists (the MD simulation), but the opposite direction is of interest.[147] The Henderson theorem states that the potential for a given RDF is unique up to a constant.[48, 49] IBI seems to converge to the RDF, but the potential is still changing even after thousands of iterations.[53] Its use is nevertheless prevalent due to its very simple “one-line” implementation. IMC on the other hand, i.e., the Newton iteration, while in theory the better choice, is not without its problems. Sampling the cross-correlations of particle counts for the IMC update is computationally expensive. Even though it is more rigorous than IBI, it has stability issues as it sometimes produces noisy potentials. Regularization, scaling of the potential update, or pre-iterations of IBI are therefore often employed.[121, 123] Despite these fixes, it seems that it is not understood why IMC has instabilities that IBI does not have.

In two previous papers, we have introduced a new type of structural coarse-graining method for solving the inverse problem which is based on integral equation theory.[75, 148] It uses the derivative of the hypernetted-chain (HNC) equation for an approximative relation between RDF and potential. The new iteration methods lead to convergence in a similar number of iterations as IMC while requiring no sampling of cross-correlation matrices. So far, our method was only available for systems of single beads. General systems with multiple beads and bonds can be treated with the Reference Interaction Site Model (RISM),[106] for which we provide the derivative in this paper. We also make use of symmetry reduction of the RISM equations that arise when multiple beads in a molecule are equivalent.[149]

Two of the significant challenges in coarse-graining are transferability across different

---

---

state points and composition as well as representability in terms of thermodynamic properties such as pressure or compressibility. In recent publications, it was argued that different additional terms are needed to achieve those. Examples are temperature-dependent potentials[150], local density-dependent potentials[86], and volume-dependent potentials[63]. At the same time, it is unclear which properties can be represented well with pair potentials alone. Many thermodynamic properties can be expressed as a function of the RDF and the pair potential and therefore be constrained using a Gauss-Newton scheme.[148, 151] In this paper, we introduce three different thermodynamic constraints and test whether they can be combined.

The paper has the following structure: In section 4.2.1 we first introduce the general scheme of the iterative methods used. In sections 4.2.2 to 4.2.5 the new general integral equation ansatz to connect structure and potential is derived. In sections 4.2.6 to 4.2.10 we introduce several modifications of the algorithm that address speed, thermodynamic consistency, and stability. In the remaining sections we apply the methods to a neon-argon mixture, hexane, and various organic solvents. All methods and algorithms used in this paper have been implemented in Python and Numpy and integrated into the VOTCA software package.[87, 88] The code resides currently in the `csg/multi-ii2` branch on <https://github.com/votca/votca>, but will be merged into the main branch soon. We also make all simulation files, RDFs, and potentials available at <https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/3492>.

## 4.2 Theory

The problem we are solving is finding a set of potentials that reproduces a given set of RDFs. The target RDFs are typically obtained from a mapped atomistic trajectory, i.e. an atomistic trajectory analyzed in terms of the CG degrees of freedom. We assume the mapping scheme that relates the position of the CG beads to the atoms (usually the center of mass) is fixed. The iterative approach is to use an initial potential guess and then make repeated updates based on the remaining difference between the current and target RDFs.

### 4.2.1 Newton and Gauss-Newton Methods

For a system with different beads, we are interested in all potentials  $u_{\alpha\beta}$  between them, where  $\alpha, \beta$  are indices for bead types. We stack all potentials to a single vector  $u = (u_{\alpha\alpha}, u_{\alpha\beta}, \dots)$ . For a system with  $n_t$  bead types there are  $n_i = \frac{n_t(n_t+1)}{2}$  interactions. Note, that each potential is itself a tabulated potential  $u_{\alpha\beta} = (u_{\alpha\beta 1}, u_{\alpha\beta 2}, \dots, u_{\alpha\beta n_{\text{cut}}})$  and  $u$  can be seen as a single one dimensional vector. The index  $n_{\text{cut}}$  equals the number of grid points up to the cut-off  $r_{\text{cut}}$ . Similarly, we define the vector  $g$  from all  $n_i$  RDFs. In

Newton's method the potentials are iteratively updated as

$$u_{k+1} = u_k - \underbrace{\mathbf{J}^{-1} \overbrace{(g_k - g_{\text{tgt}})}^{\Delta g_k}}_{\Delta u_k}. \quad (4.1)$$

The subscript  $k$  denotes iteration  $k$  and  $g_k$  are the RDFs obtained from MD with potentials  $u_k$ . Matrix  $\mathbf{J}$  is the Jacobian with the elements

$$J_{\alpha\beta i, \gamma\epsilon j} = \frac{\partial g_{\alpha\beta i}}{\partial u_{\gamma\epsilon j}}. \quad (4.2)$$

Here, the indices  $i$  and  $j$  represent the distance between the interacting beads  $\alpha, \beta$  and  $\gamma, \epsilon$ , respectively. We note that eq. (4.1) provides the explicit formulation of the potential update, but in practice it is advantageous to solve the system of linear equations without inverting the matrix.

The Jacobian  $\mathbf{J}$  is assumed to be square, i.e., the potential tables have as many data points as the RDFs used for the update. We relax this condition and optionally introduce a diagonal weighting matrix  $\mathbf{W}$  and use a Gauss-Newton algorithm, where the update is computed as

$$u_{k+1} = u_k - \underbrace{(\underbrace{\mathbf{J}_w^T \mathbf{J}_w}_{\text{generalized inv. } \mathbf{J}_w^g})^{-1} \mathbf{J}_w^T \mathbf{W}}_{\Delta u_k} \overbrace{(g_k - g_{\text{tgt}})}^{\Delta g_k} \quad (4.3)$$

with the weighted Jacobian  $\mathbf{J}_w = \mathbf{W}\mathbf{J}$ . Each RDF can now have more grid points than the potentials:  $g_{\alpha\beta} = (g_{\alpha\beta 1}, g_{\alpha\beta 2}, \dots, u_{\alpha\beta n_{\text{res}}})$  with  $n_{\text{res}} \geq n_{\text{cut}}$ . This type of update minimizes the norm

$$\|\mathbf{W}(g_{\text{tgt}} - g_k - \mathbf{J}(u_{k+1} - u_k))\|_2. \quad (4.4)$$

In other words: the potentials are updated such that the weighted difference between the target RDFs  $g_{\text{tgt}}$  and the new expected RDFs  $g_{k+1} \approx g_k + \mathbf{J}(u_{k+1} - u_k)$  is minimized. This type of formulation has several advantages. Firstly, it allows one to restrict the potential (update) to a cut-off  $r_{\text{cut}}$  while optimizing the RDF up to a longer (residuum-)cut  $r_{\text{res}}$ . Optionally, the tail can be given a larger weight in the optimization by the weighting matrix. Secondly, the scheme allows for the optimization of a selection of RDFs by changing a similar or smaller subset of potentials. This selective scheme is introduced in section 4.2.7. Thirdly, constraints can be introduced, see section 4.2.8. Before we explore those topics, we will explain how the Jacobian can be obtained.

## 4.2.2 IBI and IMC

When the derivative in eq. (4.2) is approximated by the corresponding derivative for the low density approximation

$$u = -k_B T \log(g) \quad (4.5)$$

the Jacobian becomes a diagonal matrix. One can show that[75] the resulting low density Newton scheme is numerically close to the IBI scheme[52]

$$u_{k+1} = u_k + \underbrace{k_{\text{B}}T \log \left( \frac{g_k}{g_{\text{tgt}}} \right)}_{\Delta u_k}, \quad (4.6)$$

which itself is not a Newton scheme. The low-density Newton scheme and the IBI method are local in two senses: (i) the interaction of two bead types is not influenced by the structure of a different pair of bead types and (ii) the potential update at each distance only depends on the RDF mismatch at that distance.

Inverse Monte Carlo, on the other hand, is the full Newton scheme.[50] IMC is expressed using the quantity  $S_{\alpha\beta}$ , which describes the total number of bead pairs between bead types  $\alpha$  and  $\beta$ . It is related to the RDF via  $S_{\alpha\beta i} = \frac{4\pi r_i^2 N_\alpha N_\beta}{(1+\delta_{\alpha\beta})V} g_{\alpha\beta i}$ , where  $r_i$  is the distance,  $N_\alpha$  is the number of beads of type  $\alpha$ , and  $V$  is the system volume. The factor  $(1 + \delta_{\alpha\beta})$  arises because  $S_{\alpha\beta}$  is a measure for the number of pair interactions and double-counting has to be avoided when  $\alpha = \beta$ . For a system with pair potentials, the elements of the IMC Jacobian can be computed from the cross-correlations between the particle numbers  $S_{\alpha\beta i}$  at different distances

$$\begin{aligned} A_{\alpha\beta i, \gamma\epsilon j}^{\text{IMC}} &= \frac{\partial S_{\alpha\beta i}}{\partial u_{\gamma\epsilon j}} \\ &= \frac{1}{k_{\text{B}}T} (\langle S_{\alpha\beta i} \rangle \langle S_{\gamma\epsilon j} \rangle - \langle S_{\alpha\beta i} S_{\gamma\epsilon j} \rangle). \end{aligned} \quad (4.7)$$

eq. (4.7) contains no approximations and is only limited by sampling. The matrix  $\mathbf{A}$  is the Jacobian in terms of  $S$  and  $u$ . The IMC Jacobian  $\mathbf{J}^{\text{IMC}}$  in terms of  $g$  and  $u$  is expressed as

$$J_{\alpha\beta i, \gamma\epsilon j}^{\text{IMC}} = \frac{\partial g_{\alpha\beta i}}{\partial S_{\alpha\beta i}} \frac{\partial S_{\alpha\beta i}}{\partial u_{\gamma\epsilon j}} = \frac{(1 + \delta_{\alpha\beta})V}{4\pi r_i^2 N_\alpha N_\beta} A_{\alpha\beta i, \gamma\epsilon j}^{\text{IMC}}. \quad (4.8)$$

### 4.2.3 Integral Equation Theory

Another way to obtain a connection between the potential and the RDF is integral equation theory. In particular, the Ornstein Zernike equation defines the direct correlation function  $c_{\alpha\beta}$  between beads  $\alpha$  and  $\beta$ , see section 4.2.4. Together with a closure relation, one obtains a direct but approximate relation between the potential and the RDF. Note that we use translation-invariant integral equation theory and thereby assume a homogeneous system.

We use either the hypernetted-chain closure equation[94]

$$u_{\alpha\beta} = k_{\text{B}}T(h_{\alpha\beta} - \log(g_{\alpha\beta}) - c_{\alpha\beta}) \quad (4.9)$$

---

or the Percus-Yevick closure equation

$$u_{\alpha\beta} = k_{\text{B}}T \log(1 - c_{\alpha\beta}/g_{\alpha\beta}). \quad (4.10)$$

Here,  $h_{\alpha\beta} = g_{\alpha\beta} - 1$  is the total correlation function between bead types  $\alpha$  and  $\beta$ . The direct correlation functions  $c$  can be obtained from the RDFs  $g$ , as explained in the next section. eqs. (4.9) and (4.10) can be used as an initial guess for the potential, which offers a better starting point than the potential of mean force from eq. (4.5) (often called Boltzmann inversion).[73]

For the use in a Newton's method or Gauss-Newton algorithm the Jacobian is needed. From eqs. (4.9) and (4.10) we can obtain an approximation to the inverse of the Jacobian by differentiation. The elements of the Jacobian inverse using the hypernetted-chain closure are

$$\frac{\partial u_{\alpha\beta i}}{\partial g_{\gamma\epsilon j}} = k_{\text{B}}T \left( \delta_{\alpha\gamma} \delta_{\beta\epsilon} \delta_{ij} \left( 1 - \frac{1}{g_{\alpha\beta i}} \right) - \frac{\partial c_{\alpha\beta i}}{\partial h_{\gamma\epsilon j}} \right). \quad (4.11)$$

We call the iterations based on this Jacobian HNCN or HNCNGN, depending on whether Newton or Gauss-Newton is used. The derivative  $\frac{dc}{dh}$  has to be obtained independently, see section 4.2.5. In principle, the Percus-Yevick closure can also be differentiated, but the implementation of this closure relation is less straightforward and, so far, untested. In this paper, we only consider the hypernetted-chain closure.

In eqs. (4.9) and (4.10) the pair potential is assumed to be the total pair potential between two beads. Additional potential terms, e.g., charges in the CG model, prohibit the meaningful use of integral equation methods in the given form.

In integral equation theory, the quantities  $x_{\alpha\beta}$  and  $x_{\beta\alpha}$ , where  $x$  can be  $u$ ,  $g$ ,  $h$ ,  $\omega$ , or  $c$ , are distinguished (even though they are usually equal). This is not the case outside of sections 4.2.3 to 4.2.5, where only  $x_{\alpha\beta}$  is considered. This detail has subtle implications on how to use the results from eqs. (4.9) to (4.11): To use the potential from eqs. (4.9) and (4.10) in an MD simulation,  $u_{\alpha\beta}$  and  $u_{\beta\alpha}$  are averaged. The Jacobian (inverse of eq. (4.11)) initially contains derivatives with respect to  $u_{\epsilon\gamma}$  and  $u_{\gamma\epsilon}$ . Those elements are summed up and reduced to one element. Elements that describe the derivatives of  $g_{\alpha\beta}$  and  $g_{\beta\alpha}$  are averaged, and one of them is removed, such that the Jacobian is square again.

## 4.2.4 Direct Correlation Function

To evaluate the above equations, we need the direct correlation function and its derivative. The Ornstein-Zernike (OZ) equation defines the direct correlation function. If there are any bonds between the beads, the Reference Interaction Site Model (RISM) equation has to be used, which can be seen as a generalization of the OZ equation.[94, 106] The OZ equation and the RISM equation are matrix equations for systems with different bead types. Therefore, here we introduce a new arrangement of the RDF data: a matrix  $g$  with elements  $g_{\alpha\beta}$  in row  $\alpha$  and column  $\beta$ . Each element can again be a function of the

distance  $r$ , but in this section, there will be mostly Fourier transformed variables and then each element is a function of the wave vector.

Not considering any indistinguishability between beads in a molecule, the RISM theory is written as a matrix equation

$$\hat{\mathbf{h}}' = \hat{\omega}' \hat{\mathbf{c}}' (\hat{\omega}' + \rho \hat{\mathbf{h}}'). \quad (4.12)$$

Here,  $\hat{\mathbf{h}}'$  and  $\hat{\mathbf{c}}'$  are matrices of the total and direct correlation function, respectively. The matrix  $\hat{\omega}'$  contains the intramolecular density functions. The hats denote the three-dimensional Fourier transform of the radially symmetric functions.[75] By using the Fourier transform, the convolution that takes place in real space becomes a multiplication and the matrix equation can be solved for each wave vector independently. The matrices are square with dimension  $n_{\text{td}}$ , the number of beads per molecule. For example,  $n_{\text{td}}$  is 12 for benzene in atomistic resolution. Input into the equation are the total correlation functions  $h$  and the intramolecular density functions  $\omega$ . For an initial guess, both can be obtained from the reference system ( $h_{\text{tgt}}$ ,  $\omega_{\text{tgt}}$ ).

Some beads might be equivalent, considering the bead type and symmetry of the molecule. This equivalence allows for a similarity transformation to a symmetry reduced RISM equation as derived by Bertagnolli et al.[149] The needed transformation matrix  $\mathbf{T}$  is block-diagonal with  $n_t$  blocks where  $n_t$  is the number of different bead types. Each block  $\mathbf{T}^\alpha$  is an  $n_\alpha \times n_\alpha$  matrix where  $n_\alpha$  is the number of indistinguishable beads of type  $\alpha$ . E.g.  $n_t = 2$ ,  $n_C = 6$ ,  $n_H = 6$ , for benzene in atomistic resolution. The elements of the first column of  $\mathbf{T}^\alpha$  are all equal  $n_\alpha^{-\frac{1}{2}}$ . The other columns of  $\mathbf{T}^\alpha$  have to form an orthogonal basis such that each block  $\mathbf{T}^\alpha$  is an orthonormal matrix. We obtain the transformed matrices

$$\begin{aligned} \hat{\mathbf{H}} &= \mathbf{T}^{-1} \hat{\mathbf{h}}' \mathbf{T} \\ \hat{\mathbf{\Omega}} &= \mathbf{T}^{-1} \hat{\omega}' \mathbf{T} \\ \hat{\mathbf{C}} &= \mathbf{T}^{-1} \hat{\mathbf{c}}' \mathbf{T} \\ \rho &= \mathbf{T}^{-1} \rho \mathbf{T}. \end{aligned} \quad (4.13)$$

The matrix  $\rho$  is a diagonal matrix with the densities of each bead unaffected by the similarity transformation. Inserting these definitions in eq. (4.12) and using  $\mathbf{T}^{-1} \mathbf{T} = \mathbf{T} \mathbf{T}^{-1} = \mathbf{I}_{n_{\text{td}}}$  this transformation reduces the RISM-OZ equation to

$$\hat{\mathbf{H}} = \hat{\mathbf{\Omega}} \hat{\mathbf{C}} (\hat{\mathbf{\Omega}} + \rho \hat{\mathbf{H}}). \quad (4.14)$$

In any block  $\hat{\mathbf{C}}^{\alpha\beta}$  of  $\hat{\mathbf{C}}$ , which corresponds to two bead types  $\alpha$  and  $\beta$  only one element is non-zero. The same is true for  $\hat{\mathbf{H}}$  and  $\hat{\mathbf{\Omega}}$ . Thereby the dimensionality is effectively reduced from  $n_{\text{td}}$  in eq. (4.12) to  $n_t$  in eq. (4.14) and the empty rows and columns can

be omitted (we from now on treat those matrices as  $n_t \times n_t$ ). For further details on symmetry reduction, we refer to the original paper.[149]

Note that commonly the analysis of the trajectory will average over indistinguishable beads when calculating distribution functions. Therefore one does not obtain  $\mathbf{h}'$  but  $\mathbf{h}$ . To obtain  $\mathbf{H}$ , the explicit transformation with  $\mathbf{T}$  is therefore not needed. In the appendix we give the factors needed to convert between  $\mathbf{h}$ ,  $\mathbf{c}$  and  $\mathbf{H}$ ,  $\mathbf{C}$  and clarify how  $\hat{\mathbf{\Omega}}$  is calculated.

Inverting eq. (4.14) we obtain the adapted direct correlation function from the total correlation and intramolecular density function.

$$\hat{\mathbf{C}} = \hat{\mathbf{\Omega}}^{-1} \hat{\mathbf{H}} (\hat{\mathbf{\Omega}} + \rho \hat{\mathbf{H}})^{-1}. \quad (4.15)$$

Afterwards, the direct correlation function matrix  $\hat{\mathbf{c}}$  is obtained by the inverse of the prefactor from eq. (4.44).

## 4.2.5 Direct Correlation Function Derivative

For the derivative of the HNC equation (eq. (4.11)), we need the derivative of the direct correlation function with respect to the total correlation function. When using the symmetry-adapted RISM we first have to relate the derivative to the adapted variables. Using the relation between the elements of  $\mathbf{C}$  to  $\mathbf{c}$  and  $\mathbf{H}$  to  $\mathbf{h}$  from eq. (4.44) in the appendix the following holds

$$\frac{\partial \hat{c}_{\alpha\beta}}{\partial \hat{h}_{\gamma\epsilon}} = \left( \frac{n_\gamma n_\epsilon}{n_\alpha n_\beta} \right)^{\frac{1}{2}} \frac{\partial \hat{C}_{\alpha\beta}}{\partial \hat{H}_{\gamma\epsilon}} \quad (4.16)$$

To obtain the derivative on the right hand side of eq. (4.16) we start from eq. (4.15) for  $\hat{\mathbf{C}} = \hat{\mathbf{C}}(\hat{\mathbf{H}})$  and add a perturbation  $\Delta \hat{\mathbf{H}}$  to  $\hat{\mathbf{H}}$  to define

$$\hat{\mathbf{C}}(\hat{\mathbf{H}} + \Delta \hat{\mathbf{H}}) = \hat{\mathbf{\Omega}}^{-1} (\hat{\mathbf{H}} + \Delta \hat{\mathbf{H}}) \left[ \hat{\mathbf{\Omega}} + \rho (\hat{\mathbf{H}} + \Delta \hat{\mathbf{H}}) \right]^{-1}. \quad (4.17)$$

This expression is transformed in several steps, writing  $\hat{\Omega} + \rho\hat{\mathbf{H}}$  as  $\mathbf{Z}$  for brevity and ignoring higher order terms

$$\begin{aligned}
& \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}}) \left[ \mathbf{Z} + \rho\Delta\hat{\mathbf{H}} \right]^{-1} \\
&= \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}}) \left[ \mathbf{Z}(\mathbf{I} + \mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}}) \right]^{-1} \\
&= \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}})(\mathbf{I} + \mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}})^{-1}\mathbf{Z}^{-1} \\
&= \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}})(\mathbf{I} - \mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}})\mathbf{Z}^{-1} \\
&= \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}})(\mathbf{Z}^{-1} - \mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1}) \\
&= \hat{\Omega}^{-1}\hat{\mathbf{H}}\mathbf{Z}^{-1} + \hat{\Omega}^{-1}\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1} \\
&\quad - \hat{\Omega}^{-1}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}})\mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1} \\
&= \hat{\mathbf{C}} + \hat{\Omega}^{-1}\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1} - \hat{\Omega}^{-1}\hat{\mathbf{H}}\mathbf{Z}^{-1}\rho\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1} \\
&= \hat{\mathbf{C}} + \hat{\Omega}^{-1}(\mathbf{I} - \hat{\mathbf{H}}\mathbf{Z}^{-1}\rho)\Delta\hat{\mathbf{H}}\mathbf{Z}^{-1}
\end{aligned}$$

and finally

$$\hat{\mathbf{C}}(\hat{\mathbf{H}} + \Delta\hat{\mathbf{H}}) - \hat{\mathbf{C}}(\hat{\mathbf{H}}) = \hat{\Omega}^{-1}(\mathbf{I} - \hat{\mathbf{H}}(\hat{\Omega} + \rho\hat{\mathbf{H}})^{-1}\rho) \cdot \Delta\hat{\mathbf{H}}(\hat{\Omega} + \rho\hat{\mathbf{H}})^{-1}. \quad (4.18)$$

We apply vectorization to eq. (4.18), which transforms a matrix into a long vector going through the elements column-wise. Using the identity  $\text{vec}(\mathbf{U}\mathbf{V}\mathbf{W}) = (\mathbf{W}^T \otimes \mathbf{U}) \text{vec}(\mathbf{V})$ , where  $\otimes$  is the Kronecker product, we eventually arrive at the derivative

$$\frac{d \text{vec}(\hat{\mathbf{C}})}{d \text{vec}(\hat{\mathbf{H}})} = (\hat{\Omega} + \rho\hat{\mathbf{H}})^{-1T} \otimes \left[ \hat{\Omega}^{-1} \left( \mathbf{I} - \hat{\mathbf{H}}(\hat{\Omega} + \rho\hat{\mathbf{H}})^{-1}\rho \right) \right] \quad (4.19)$$

This derivative reduces to Eq. (9) from our previous paper[148] for the single bead case. The distributions  $\mathbf{H}$  and  $\Omega$  are obtained from the MD simulation with the current potential, except when the technique described in the next section is used.

Using eqs. (4.16) and (4.19) we obtain the derivatives  $\frac{\partial c_{\alpha\beta}}{\partial h_{\gamma\epsilon}}$ , which is a numeric value for each wave vector. To obtain the same derivative in real space, as needed for eq. (4.11), we use

$$\frac{\partial c_{\alpha\beta}}{\partial h_{\gamma\epsilon}} = \mathcal{F}^{-1} \text{diag} \left( \frac{\partial \hat{c}_{\alpha\beta}}{\partial \hat{h}_{\gamma\epsilon}} \right) \mathcal{F}, \quad (4.20)$$

where  $\mathcal{F}$  is the Fourier matrix. The derivative  $\frac{\partial c_{\alpha\beta}}{\partial h_{\gamma\epsilon}}$  is now a matrix with elements  $\frac{\partial c_{\alpha\beta i}}{\partial h_{\gamma\epsilon j}}$ .

---

## 4.2.6 One-Time Jacobian Calculation

The IMC and HNC methods are more expensive per iteration than IBI because the Jacobian matrix has to be calculated. To make those methods faster, one can attempt to calculate it only once.

The averages in eq. (4.7) are evaluated in the system with the current potential guess. A computation-intensive sampling of long MD trajectories has to be carried out at every iteration. A simple shortcut is to only calculate it once at the target using the trajectory of the reference simulation, provided the latter is available

$$A_{\alpha\beta i, \gamma\epsilon j}^{\text{t-IMC}} = \frac{1}{k_{\text{B}}T} \left( \langle S_{\alpha\beta i}^{\text{tgt}} \rangle \langle S_{\gamma\epsilon j}^{\text{tgt}} \rangle - \langle S_{\alpha\beta i}^{\text{tgt}} S_{\gamma\epsilon j}^{\text{tgt}} \rangle \right), \quad (4.21)$$

where the brackets denote in this case an average over the fine-grained trajectory. eq. (4.21) contains an approximation to the Jacobian, because the IMC formalism assumes only pairwise interactions. The configurations in the mapped atomistic trajectory are, however, determined by the multibody potential of mean force, which is only approximately decomposable in pair potentials. It is also important that the sampling of eq. (4.21) is evaluated in the NVT ensemble, not in the NpT ensemble. The reason is that under NpT conditions, the barostat changes the probabilities for multiple interactions being at a certain distance by scaling the box volume. We prepend a “t” and write t-IMC to denote that the IMC matrix was calculated at the target configurations.

A similar shortcut can be taken for the HNC Jacobian as described in eq. (4.11). In particular, we have implemented the option to calculate  $\frac{\partial c}{\partial h}$  only once from the target distributions. As described in our last paper, this derivative contains artifacts if not calculated from an RDF with double length.[148] When  $\frac{\partial c}{\partial h}$  is calculated only once at the very beginning, then it is sufficient to calculate the RDF on the short length of the potential in all subsequent iterations. This leads to a considerable performance improvement since the computational cost of calculating the RDF grows rapidly with the maximum pair distance. We prepend a “t” and write t-HNCN or t-HNCGN if the derivative  $\frac{\partial c}{\partial h}$  is calculated only once from the target RDF.

## 4.2.7 Selective Optimization

In section 4.2.1 we explained that the Gauss-Newton method, by default, minimizes the norm  $\|\mathbf{W}(g_{\text{tgt}} - g_k - \mathbf{J}(u_{k+1} - u_k))\|_2$ . In this formulation  $u$  and  $g$  encompass all  $n_i$  non-bonded interactions. In some situations, it is desirable only to update some potentials and ignore the match of some RDFs. To achieve this, eq. (4.3) can be adapted. All potentials that should not be updated have their respective columns in the Jacobian removed. All RDFs that shall not be considered targets have their respective rows removed from the Jacobian and the residuum vector. For the HNC methods, the full Jacobian always has to

be constructed first and then reduced. For IMC, interactions can be left out from the start when generating the IMC matrix.

## 4.2.8 Constraints

When using the Gauss-Newton algorithm, each iteration  $k$  consists of a linear least squares problem, because eq. (4.4) can be rewritten as a minimization of

$$\|\mathbf{W}(\mathbf{J}\Delta u_k + \Delta g_k)\|_2 \quad (4.22)$$

with  $\Delta g_k = g_k - g_{\text{tgt}}$ , which is solved for  $\Delta u_k$ . The potential update can be required to meet certain criteria by constraining the solution of the least squares problem to fulfill

$$\mathbf{K}\Delta u_k = d. \quad (4.23)$$

The number of rows in the matrix  $\mathbf{K}$  and vector  $d$  equals the number of linear constraints. Generally, for any function  $\phi(r, g, u)$  of the pair distance, RDF, and pair potential a constraint can be derived. To show this,  $\phi$  is expanded up to the first order

$$\phi(r, g + \Delta g, u + \Delta u) = \phi(r, g, u) + \frac{\partial \phi}{\partial g} \Delta g + \frac{\partial \phi}{\partial u} \Delta u. \quad (4.24)$$

Here,  $\phi(r, g, u)$  is the value of the function in the current iteration, i.e.  $\phi_k$ . If we define  $\phi(r, g + \Delta g, u + \Delta u)$  as  $\phi_{\text{tgt}}$ , reorder eq. (4.24), and use  $\Delta g \approx -\mathbf{J}\Delta u$  we obtain

$$\underbrace{\left( \frac{\partial \phi}{\partial g} \mathbf{J} - \frac{\partial \phi}{\partial u} \right)}_{\mathbf{K}_\phi} \Delta u = \underbrace{\phi_k - \phi_{\text{tgt}}}_{d_\phi}. \quad (4.25)$$

This form gives us a row in  $\mathbf{K}$  and an element in  $d$ .

From this general formulation, we derive the pressure constraint for multi-bead systems. The function  $\phi$  is in that case the virial pressure  $p$  with the integral over the  $n_{\text{cut}}$  grid points of  $r$  in discrete form

$$p = \rho k_B T + \sum_{\alpha \leq \beta} \frac{(2 - \delta_{\alpha\beta}) \rho_\alpha \rho_\beta}{6} \times \sum_{i'}^{n_{\text{cut}}} r_{i'} f_{\alpha\beta}(r_{i'}) g_{\alpha\beta}(r_{i'}) 4\pi r_{i'}^2 \Delta r. \quad (4.26)$$

The factor  $2 - \delta_{\alpha\beta}$  arises because the double sum over  $\beta$  and  $\alpha$  is usually unrestricted in the virial. The index  $i'$  runs over grid points shifted by half the grid spacing  $\Delta r$  from the usual grid. The values of the pair distance, RDF, and force at those in-between grid points are defined by  $r_{i'} = \frac{r_i + r_{i+1}}{2}$ ,  $g_{\alpha\beta}(r_{i'}) = \frac{g_{\alpha\beta i} + g_{\alpha\beta i+1}}{2}$ , and  $f_{\alpha\beta}(r_{i'}) = -\frac{u_{\alpha\beta i+1} - u_{\alpha\beta i}}{\Delta r}$ , respectively. For the pressure expansion we neglect  $\frac{\partial p}{\partial g}$  in eq. (4.25). This is a good

approximation for all but the first few iterations, after which the changes in the RDF are usually much smaller than the changes in the potential. The elements of row  $K_p$  in  $\mathbf{K}$  are obtained as

$$\begin{aligned} K_{p,\alpha\beta i} &= -\frac{\partial p}{\partial u_{\alpha\beta i}} \\ &= -\frac{2\pi(2 - \delta_{\alpha\beta})\rho_\alpha\rho_\beta}{3} \\ &\quad \times (g_{\alpha\beta i}r_i^3 - g_{\alpha\beta i+1}r_{i+1}^3). \end{aligned} \quad (4.27)$$

The value of the element of  $d$  is given by  $d_p = p_k - p_{\text{tgt}}$ . This formulation for the pressure constraint is different from that in our previous papers[75, 148], where the Gauss-Newton algorithm would update the force instead of the potential. Another difference is that in the previous papers,  $p_{\text{tgt}}$  was used in the constrain formula while we here assume  $\Delta g$  is small and therefore use  $g_k$ .

Next, we introduce a constraint that aims to match the Kirkwood-Buff integral (KBI) of the target RDF. The KBI between two species  $\alpha$  and  $\beta$  in a homogeneous system can be written in discrete form as

$$G^{\alpha\beta} = 4\pi \sum_i (g_{\alpha\beta i} - 1)r_i^2 \Delta r. \quad (4.28)$$

Here,  $i$  is the index running over distances which in theory should go to infinity and in practice goes up to  $n_{\text{res}}$ . Applying eq. (4.25) on eq. (4.28) we obtain

$$K_{G^{\alpha\beta}, \gamma\epsilon j} = 4\pi \Delta r \sum_i^{n_{\text{res}}} r_i^2 J_{\alpha\beta i, \gamma\epsilon j} \quad (4.29)$$

and

$$d_{G^{\alpha\beta}} = G_k^{\alpha\beta} - G_{\text{tgt}}^{\alpha\beta}. \quad (4.30)$$

$G_k^{\alpha\beta}$  and  $G_{\text{tgt}}^{\alpha\beta}$  are obtained from the current and target RDF and eq. (4.28). In a system with multiple interactions, the constraint can be applied to multiple or all interactions.

The intermolecular (cohesive) potential energy of a system determines its vaporization enthalpy. In a system of only bonded and pair interactions, it can be expressed as

$$PE^{\text{inter}} = V \sum_{\alpha \leq \beta} \frac{(2 - \delta_{\alpha\beta})\rho_\alpha\rho_\beta}{2} \times \sum_i u_{\alpha\beta}(r_i)g_{\alpha\beta}(r_i)4\pi r_i^2 \Delta r. \quad (4.31)$$

We note, that eq. (4.31) does not explicitly exclude intramolecular pair potentials for large molecules. The formulation here given was only tested for molecules without intramolecular pair potential interactions. We use eq. (4.25), ignore  $\frac{\partial PE^{\text{inter}}}{\partial g}$  as for the pressure

constraint, and constrain  $PE^{\text{inter}}/N$  to obtain a system size independent formulation

$$\begin{aligned} K_{PE,\alpha\beta i} &= \frac{\partial PE^{\text{inter}}/N}{\partial u_{\alpha\beta i}} \\ &= 2\pi\Delta r \frac{V}{N} (2 - \delta_{\alpha\beta}) \rho_{\alpha} \rho_{\beta} g_{\alpha\beta i} r_i^2 \end{aligned} \quad (4.32)$$

and

$$d_{PE,\alpha\beta i} = PE_k^{\text{inter}} - PE_{\text{tgt}}^{\text{inter}}. \quad (4.33)$$

We prepend *p*-, *PE*-, or *KBI*- to the method name to denote that the pressure, intermolecular potential energy, or Kirkwood-Buff integrals of the atomistic system are set as constraints to the potential update, respectively. The *KBI*- prefix denotes that all non-bonded interactions have been constrained, so this is not a single constraint but multiple. A surface tension constrained as described for IMC[123] was not implemented, but we note that our code is easily extensible.

To solve the least squares problem with constraints, we use a direct elimination algorithm as described in [78].

#### 4.2.9 RDF Onset Extrapolation

Due to the Pauli exclusion principle, atoms never get to be in the same position. This leads to all atomistic RDFs starting with zero. Coarse-grained beads behave similarly if the mapping scheme does not aggregate too many atoms in one bead. Therefore, there is usually an onset region where the RDF goes from zero to nonzero values. This region suffers from poor sampling as few beads ever get to come this close together. The sampling problem becomes visible in a logarithmic plot of the RDF, e.g., in fig. 4.1. There is visible noise at low values even with long sampling, and below some threshold, all values are zero. The noise and discontinuity can lead to critical problems for coarse-graining algorithms. For example, if the initial potential guess is the potential of mean force (PMF) (same as eq. (4.5)) then the repulsive part of the potential can become very noisy or steep, causing instabilities in the MD simulation. Different problems arise for potential updates and will be discussed in section 4.2.10.

We assume that in the onset region the RDF follows the asymptotics

$$g(r) \propto e^{-\frac{u(r)}{k_{\text{B}}T}} \quad (4.34)$$

of the PMF. eq. (4.34) implies that, up to an additive constant, the PMF matches the pair potential in the onset region if the RDF is perfectly sampled. To extrapolate, we assume that the potential in the repulsive region can be described with an exponential decay. We therefore propose to take  $-\log(g)$ , fit it with  $f = ae^{-bx} + c$  in the well-sampled region, extrapolate into the poorly sampled region, and then take  $e^{-f}$  to obtain the

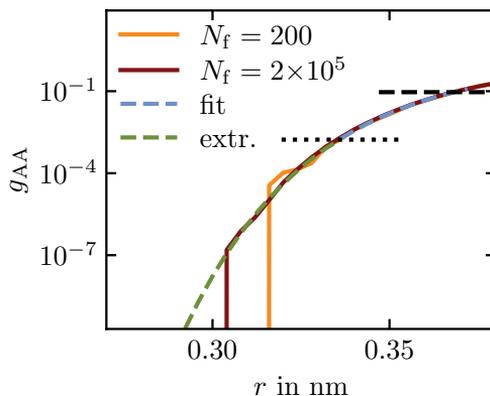


Figure 4.1: Two A-A RDFs with different sampling were obtained from an OPLS/AA hexane liquid trajectory mapped with a three-bead scheme (see section 4.3.2). The system has 3000 molecules, and  $N_f$  is the number of trajectory frames analyzed. Also shown is an extrapolation which was obtained from fitting the potential of mean force of the  $N_f = 200$  RDF in the region between the dotted and dashed horizontal lines at 0.001 and 0.1 with an exponential function.

extrapolated RDF. The region for the fit has to be at the onset of the RDF, where eq. (4.34) holds. At the same time, the values have to be reasonably well-sampled. We made good experiences with taking the region where the RDF has values between 0.0001 and 0.1 and extrapolating all values that are below 0.0001. In fig. 4.1, we show an example for an extrapolation of a poorly sampled RDF. It extrapolates the RDF smoothly to very small values. We find it more consistent to extrapolate the RDFs before deriving potentials rather than extrapolating the potentials later. It also simplifies the usage of constraints in the potential update, which get violated when altering the potential after the Gauss-Newton update.

#### 4.2.10 RDF Onset Newton Instabilities

Even with the RDF onset extrapolated, as described in the last section, there is a subtle detail that often leads to artifacts in the updated potential obtained from Newton's methods. The following observation can often be made when using IMC, independent of the sampling. When the current RDF has lower values at the onset region compared to the target RDF, the Newton update will overshoot to a very negative potential. In practice, one will observe deep dips in the repulsive region of the updated potential and small peaks in the RDF onset region. We illustrate the cause of this effect in fig. 4.2,

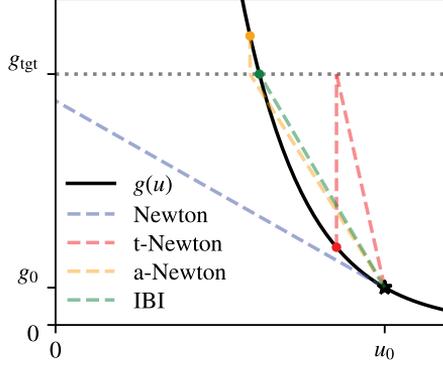


Figure 4.2: The analytical relation of eq. (4.34) between a single value of the RDF and the potential value at the same distance in the onset region. Starting from potential value  $u_0$ , we show different update schemes and their value after one iteration. The stars shows the starting point and the colored dots show  $u$  and  $g$  after the update.

where we show  $g$  as a function of  $u$  at a fixed distance assuming the analytical form from eq. (4.34). The overshoot from  $u_0$  to negative  $u$  with Newton's method happens because the derivative at  $u_0$  is very small. In other words, in the onset region, it happens very quickly that current and target RDF have values that are small but orders of magnitude different (e.g.,  $10^{-3}$  and  $10^{-5}$ ) and given the non-linear relation between  $g(r)$  and  $u(r)$  a simple Newton update easily leads to artifacts. When using IMC, there is little sampling in the onset region leading to additional noise. The t-Newton's method in fig. 4.2 gets the derivative from  $g_{\text{tgt}}$  (compare section 4.2.6) and leads potential update which is too small. Note that for  $g_0 > g_{\text{tgt}}$  (not shown) the behaviour of Newton and t-Newton switches. Interestingly, IBI does not have this problem. Instead, this is the region, where the IBI method works best, as can be seen in fig. 4.2. This is to be expected, since an IBI update has the form

$$u_{k+1}^{\text{IBI}} = u_k + k_{\text{B}}T \log \frac{g_k}{g_{\text{tgt}}} \quad (4.35)$$

and if we insert first  $g = g_k$ ,  $u = u_k$  and second  $g = g_{k+1}^{\text{IBI}}$ ,  $u = u_{k+1}^{\text{IBI}}$  into eq. (4.34), divide the two resulting equations, and apply eq. (4.35) we obtain

$$\frac{g_k(r)}{g_{k+1}^{\text{IBI}}(r)} = e^{-\frac{u_k(r)}{k_{\text{B}}T} + \frac{u_k(r) + k_{\text{B}}T \log \frac{g_k(r)}{g_{\text{tgt}}(r)}}{k_{\text{B}}T}} = \frac{g_k(r)}{g_{\text{tgt}}(r)} \quad (4.36)$$

$$g_{k+1}^{\text{IBI}}(r) = g_{\text{tgt}}(r).$$

Therefore, IBI will result in the exact update needed to match the RDF in one iteration under the condition of eq. (4.34). We believe that the stability of IBI played a considerable role in its success in coarse-graining. IBI will always converge well in the onset region and prevent crashes caused by unphysical potentials.

To retain some of the IBI stability with Newton's method, we can modify the Jacobian in the onset region

$$J_{\alpha\beta i, \alpha\beta i}^{\text{a-Newton}} = -\frac{1}{k_{\text{B}}T} \bar{g}(r). \quad (4.37)$$

Here,  $\bar{g} = (g_{\text{tgt}} + g_k)/2$  is the average of the current and the target RDF. This option, shown in fig. 4.2 as a-Newton (a for average), converges satisfactorily in the onset region. Note that we do not use the IBI update, since it is not a Newton scheme and we are interested in a unified formulation with a Jacobian. We have implemented a modification for IMC by changing the Jacobian according to

$$J_{\alpha\beta i, \gamma\epsilon j}^{\text{IMC, impr.}} = \begin{cases} \underbrace{\delta_{\alpha\gamma} \delta_{\beta\epsilon} \delta_{ij} \frac{-1}{k_{\text{B}}T} \bar{g}_{\alpha\beta}}_a, & \text{if } g^\dagger < t_1 \\ \underbrace{J_{\alpha\beta i, \gamma\epsilon j}^{\text{IMC}}}_b, & \text{if } g^\dagger > t_2 \\ (1-x)a + xb, & \text{otherwise,} \end{cases} \quad (4.38)$$

with  $x = \frac{\sqrt{g^\dagger - t_1}}{\sqrt{t_2 - t_1}}$

and  $g^\dagger = \begin{cases} g_{\alpha\beta i}, & \text{if } j \geq i \\ g_{\gamma\epsilon j}, & \text{otherwise.} \end{cases}$

This formulation is essentially just an interpolation between the diagonal Jacobian from eq. (4.37) and the unmodified IMC Jacobian. Start point and end point of the interpolation are determined by two threshold values  $t_1$  and  $t_2$  that compare against the value of the RDF. The square root was chosen for the interpolation to compensate for the quick growth of the RDF in the onset region. In fig. 4.3 we show the effect of this modification on the potential update. Without the improvement of the IMC Jacobian, the new potential has a noisy region that can lead to instabilities in an MD simulation. In our experience, if one uses tiny MD time steps and IMC without modification, small peaks in the RDF onset region will show, which do not go away. In previous studies, IBI was often used before IMC, which makes sense due to IBI's good convergence behavior at the onset region. We have in the past used IMC and IBI alternately to get rid of those artifacts when a single IBI pre-run has not worked.[152] With our improvement to the IMC matrix, the IMC method can be used directly in most cases without any pre-run.

For the HNC methods we modify eq. (4.11) to use the average between the target

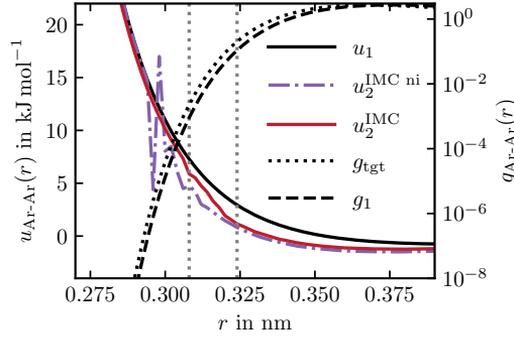


Figure 4.3: The Ar-Ar potential after an IMC update with the non-improved (ni) and improved IMC Jacobian. Also shown is the potential  $u_1$  before the update which is based on the HNC initial guess. To the right of the graph is the scale for the target and current RDF, which are plotted in a logarithmic scale. Vertical grey dotted lines mark the region in which the IMC matrix has been interpolated.

and current RDF in the diagonal term

$$\frac{\partial u_{\alpha\beta i}}{\partial g_{\gamma\epsilon j}} = k_B T \left( \delta_{\alpha\gamma} \delta_{\beta\epsilon} \delta_{ij} \left( 1 - \frac{1}{\bar{g}_{\alpha\beta i}} \right) - \frac{\partial c_{\alpha\beta i}}{\partial h_{\gamma\epsilon j}} \right). \quad (4.39)$$

Because the diagonal term dominates the derivative  $\frac{\partial c}{\partial h}$  in the onset region, this formulation leads to good convergence.

In fig. 4.4 the general process of HNCNGN as implemented in VOTCA is shown. The graph encompasses some of the topics discussed in section 4.2. It also explicitly mentions the IBI update on intramolecular interactions (bonds, angles, dihedrals). We write, without further details,  $p^{\text{intra}}$  and  $u^{\text{intra}}$  for the intramolecular probability density distributions and potentials. The formula for the IBI update of bond, angles, and dihedrals is equivalent to eq. (4.6).

### 4.3 Simulation Details

We simulate two different systems to test the algorithms that were described above. The first is a liquid 1:1 neon-argon mixture at 100 K and 1000 bar. The second is liquid n-hexane at 20 °C and 1 bar. To further test our methods and demonstrate their performance at scale, we simulate nine solvents and all their 36 mixtures at  $x = 0.5$ . The nine solvents are ethanol, 2-butanol, tert-butanol, ethylene glycol, acetone, pyridine, 1,4-dioxane, diethyl ether, and ethyl acetate. They are also considered at conditions of 20 °C and 1 bar



---

---

with a time constant of 10 ps and is simulated for  $2 \times 10^6$  steps. For the NVT production run, the average box volume from the NpT equilibration is used. Trajectory frames are saved every 10 steps.

For neon and argon the Lennard-Jones parameters are  $\sigma_{\text{Ne}} = 2.78 \text{ \AA}$ ,  $\epsilon_{\text{Ne}} = 0.28869 \text{ kJ mol}^{-1}$ ,  $\sigma_{\text{Ar}} = 3.401 \text{ \AA}$ , and  $\epsilon_{\text{Ar}} = 0.978628 \text{ kJ mol}^{-1}$ . [154, 155] Force field parameters for the organic molecules were generated with LigParGen. [138] This web service creates OPLS force fields with charge models using localized bond-charge corrections. [156, 157] Charges were slightly adjusted to account for the symmetry and the zero net-charge of the molecules. The geometric average is used to obtain  $\sigma$  and  $\epsilon$  for Lennard-Jones interactions between different atoms. Lennard-Jones potentials are shifted upwards to yield zero at the cut-off. All systems are simulated with a total of 3000 molecules in a cubic box with periodic boundary conditions. The fine-grained molecular dynamics simulations are performed with Gromacs 2021.5. [139]

### 4.3.2 Coarse-Graining

For the neon-argon mixture, the CG beads are positioned at the atoms (unity mapping). Thereby, we can test if the various methods retain the atomistic pair potentials. [54] The hexane molecules are mapped to an ABA representation where B represents the two central carbon atoms and the attached hydrogen atoms, and A beads the ethyl groups at either side. The beads are positioned at the center of mass of the atoms they represent. It contains two bonds and an angle potential as well as two identical beads and is therefore well suited to test the symmetry-adapted RISM methods based on section 4.2.4. Non-bonded interaction potentials within the ABA representation are excluded. The nine solvent molecules are mapped to single beads. The trajectory of the fine-grained production is mapped according to those mapping schemes, and the RDFs and (for hexane) bond and angle distributions are calculated. For hexane, the intramolecular density functions  $\omega$  are also calculated. For the t-IMC method (see section 4.2.6) the IMC matrix is calculated at this stage.

After the definition of the mapping scheme, the first step of coarse-graining is the initial potential guess. We mainly use the hypernetted-chain equation (eq. (4.9)) which gives a very good initial guess. [73, 148]. For its evaluation, the maximum available length of the RDF is used, which can be found in table 4.1. The initial guess is then cut at the cut-off and shifted such that it is zero beyond. The HNC initial guess is always calculated with eq. (4.9) using the whole RDF. If a t-HNC method is used (see section 4.2.6), at this stage also the derivative  $\frac{\partial c}{\partial h}$  is calculated.

At each iteration of the coarse-graining procedure, a CG MD simulation was performed with the current potential. The time step is 5 fs for the neon-argon mixture, 1.5 fs for hexane and 2 fs for all solvents. Simulations are performed in NVT using a Langevin dynamics integrator with a friction constant of  $2 \text{ ps}^{-1}$ . The density is chosen to match the density of the fine-grained production run. The CG systems have the same number of

Table 4.1: The maximum RDF length, potential cut-off  $r_{\text{cut}}$ , residuum cut-off  $r_{\text{res}}$ , and RDF grid spacing  $\Delta r$  used in the coarse-graining of different systems. The residuum length  $r_{\text{res}}$  is only used in Gauss-Newton methods with constraints. All lengths are in nm.

system	max. length	cut-off $r_{\text{cut}}$	residuum-cut $r_{\text{res}}$	$\Delta r$
neon-argon	2.4	0.9	–	0.002
hexane	4.2	1.2	1.6	0.004
solvents	3.2	1.2	1.6	0.008

molecules as the fine-grained systems. We run the CG-MD for  $2 \times 10^6$  steps for constrained Gauss-Newton methods and IMC and  $2 \times 10^5$  steps for all other methods. IMC needs more frames per iteration for the sampling of the IMC matrix. The constrained Gauss-Newton methods are run longer to report well-converged properties, but we also successfully tested them with only  $2 \times 10^5$  steps. The trajectory is saved every 10 steps, and from it, the RDFs, intramolecular density functions, bond and angle distributions, and optionally the IMC matrix and further ensemble properties are calculated. All RDFs are extrapolated in the onset region with the method described in section 4.2.9 with thresholds 0.001 and 0.1. Because CG models sample their available phase space much faster, it is sufficient to discard 100 ps from the MD run as equilibration.

Each iteration ends with the calculation of the potential update. For the IMC update, the IMC Jacobian was modified as described in section 4.2.10 with the interpolation region between the RDF values of 0.001 and 0.1. For the HNC methods, eq. (4.39) is used throughout the paper. All Gauss-Newton runs used  $r^2/(g_{\text{tgt}} + 10^{-30})$  as a diagonal on the weighting matrix  $\mathbf{W}$  to put weight onto the onset and the tail of the RDF. VOTCA has many options for post-processing of the potential (e.g., smoothing, shifting, extrapolation), but in this work, we did not use them not to convolute their effects with the effects of the CG methods. One exception is the bond and angle potentials in the hexane system, which were extrapolated in unsampled regions after each potential update. For the MD runs, the tabulated potentials are interpolated to match the fixed grid size of GROMACS, which is 0.002 nm.

All coarse-graining tasks are handled and performed by VOTCA. The Gauss-Newton scheme with constraints, the RDF onset improvement, and the integral equation methods have been integrated. The CG molecular dynamics simulations within VOTCA are performed with Gromacs 2019.6 (newer versions do not support tabulated potentials).[139]

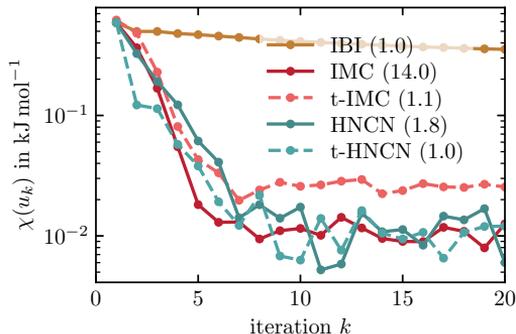


Figure 4.5: Potential-convergence behavior of different methods applied on the neon-argon mixture system. The number in brackets is the time in minutes per iteration on a dual-socket AMD EPYC 7413 machine.

## 4.4 Results

### 4.4.1 Neon-Argon

As explained before, the neon-argon system serves as a test case for the methods in this paper, as the true pair potential  $u_{\text{ref}}$  is known. In fig. 4.5 we show the potential convergence of five different methods measured by the quantity

$$\chi(u) = \sum_{\alpha \leq \beta} \sqrt{\frac{1}{r_{\text{cut}}} \int_0^{r_{\text{cut}}} g_{\alpha\beta}^{\text{tgt}} (u_{\alpha\beta} - u_{\alpha\beta}^{\text{ref}})^2 dr}. \quad (4.40)$$

The measure  $\chi(u)$  is a sum over all non-bonded interactions and uses the target RDF as a weighting to only measure potential differences in sampled regions. All methods start with the same potential guess and, therefore, the same  $\chi(u_1)$ . With IBI, the potential converges very slowly and does not reach a plateau within 20 iterations. This finding is connected to the underlying low-density approximations of the method as explained in sections 4.2.2 and 4.2.6: The other four methods, IMC, t-IMC, HNCN, and t-HNCN, reach convergence within 10 iterations. Of those, the computational cost per iteration is lowest for t-IMC and t-HNCN, as was explained in section 4.2.6. The height of the plateau is about a factor of two higher for t-IMC. This slightly worse convergence is caused by noise in the IMC matrix, which was calculated only once from  $2 \times 10^5$  atomistic frames. In contrast to IMC, t-IMC always uses the same matrix and any noise within will therefore persist across iterations. In fig. 4.6 we show the potentials that result from the different methods from fig. 4.5. All methods but IBI have converged to the expected potentials after 20 iterations. The potentials of t-IMC potentials have slightly too shallow minima. Note

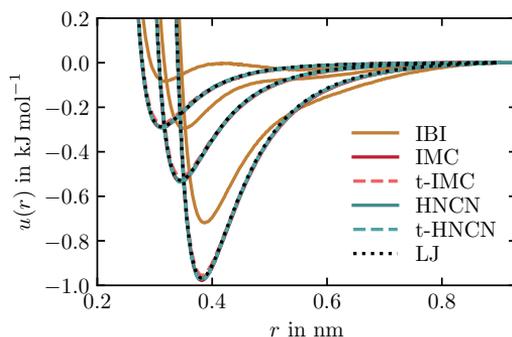


Figure 4.6: Ne-Ne, Ar-Ne, and Ar-Ar pair potential (In order of potential minimum position) after 20 iterations of different methods applied to the neon-argon mixture system. Shown as a dotted line is the LJ potential of the reference system.

that all potentials produce very similar RDFs (not shown) that are hard to distinguish by the bare eye. Our findings confirm that the HNC methods work well for mixtures. Also, it is demonstrated that by reusing the Jacobian in t-HNCN and t-IMC, coarse-graining can be sped up significantly. There is a small cost in accuracy with t-IMC, while t-HNCN is both accurate and fast.

#### 4.4.2 Hexane

Before testing convergence on the hexane system, we compare the IMC and HNC Jacobian in fig. 4.7. We find the IMC Jacobian to be much noisier even though a factor of 10 more frames were sampled for its generation. In general, the two Jacobians have similar structures and intensities. One interesting distinction is the structure of the blocks that are not on the diagonal. There we find that the HNC Jacobian shows a pattern of parallel lines. In the IMC Jacobian, those lines also exist and have similar intensity but are not strictly straight and parallel. This difference is caused by the approximate nature of the RISM equation: It only accounts for the average intramolecular distances but has no information on how they change dynamically when two molecules get close to each other. IMC has that information because it samples the Jacobian from the trajectory.

For coarse-graining hexane, potentials are needed for the angle and the two identical bonds in the CG representation. We make an initial guess by Boltzmann-inverting their target distribution functions. To further refine those bonded potentials, we tried to make IBI updates on them parallel to the update on the non-bonded potentials at every iteration. This, however, led to neither the bonded nor the non-bonded distributions converging, as shown in fig. 4.8. The problem is that the non-bonded distribution depends to some degree on the bonded potentials and vice versa. However, this connection is not reflected

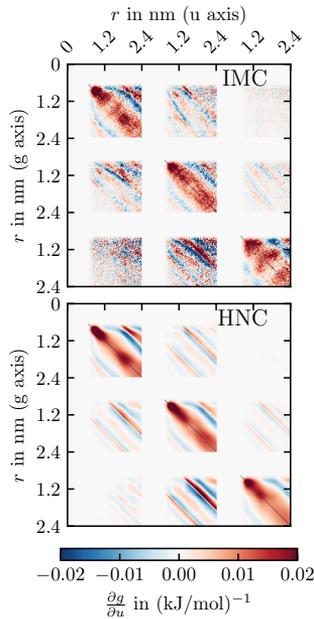


Figure 4.7: The IMC and HNC Jacobian obtained from simulation with the initial guess (for iteration 1). The rows and columns of the visible blocks are related to the interactions A-A, A-B, and B-B.

in the potential update method if both are updated independently. What happens is that both interactions are updated as if the other type of potential remains unchanged, which results in an overshoot. This leads to the type of oscillatory behavior that is seen in fig. 4.8. There are two possible ways to solve this issue. The first would be to derive an integrated update scheme that considers both bonded and non-bonded interactions. This ansatz is possible with the IMC approach but not implemented in VOTCA.[158] For an HNC-like scheme, one would need the RISM theory to predict bonded distribution functions and such an equation does, to our knowledge, not exist. A second way is to update non-bonded and bonded interactions alternately. We do this for the hexane system in the rest of the paper and find it to perfectly match the bond and angle distributions after a few iterations.

In fig. 4.9 we see the RDF convergence of different methods applied to hexane. The convergence metric is

$$\chi(g) = \sum_{\alpha \leq \beta} \sqrt{\frac{1}{r_{\text{cut}}} \int_0^{r_{\text{cut}}} (g_{\alpha\beta} - g_{\alpha\beta}^{\text{tgt}})^2 dr}. \quad (4.41)$$

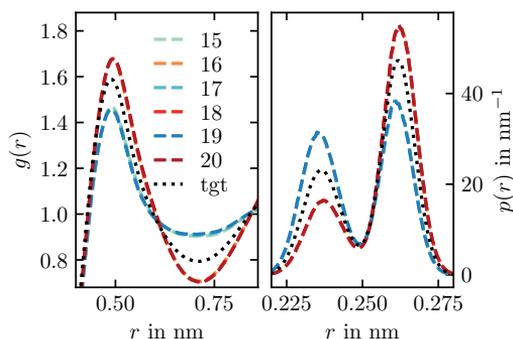


Figure 4.8: The A-A RDF and A-B bond distribution of CG hexane for six iterations of HNCN with simultaneous IBI of bonded potential. The numbers in the legend indicate the iteration  $k$ .

We find that IBI converges slowest and does not reach a plateau within 40 potential updates (20 bonded, 20 non-bonded). The other two methods converge faster, where IMC is converged after 12 iterations and HNCN after 25. While for neon-argon the convergence of the two methods was similar, here, HNC is slower. The cause behind the slower convergence is the RISM-OZ equation being a less good approximation for the potential-RDF relation, especially when the molecule is not stiff.[94] The approximations in RISM-OZ also cause the differences in the Jacobians shown in fig. 4.7. Besides slower convergence, we find that HNCN converges to a higher plateau than IMC. In principle, one would expect both methods to converge to the same potential due to the Henderson theorem. The difference, therefore, has to come from IMC having a better sampling of the RDF since MD is run ten times longer per iteration. To confirm this hypothesis, the HNCN iterations were rerun with the same number of MD time steps as used for IMC. Also shown in fig. 4.9, this confirms that HNCN converges to a similar (or even better) precision as IMC with the same amount of sampling. The results from the accelerated t-IMC and t-HNCN methods have similar convergence speeds (iterations to reach plateau) as their un-accelerated counterparts. The accuracy of t-IMC and t-HNCN, however, is reduced compared to IMC and HNCN, respectively. In fig. 4.10 we show the potential of the last iteration of the same methods from fig. 4.9. It shows that all methods except IBI find very similar potentials within 40 iterations. The converged potentials fluctuate slightly between iterations which explains the small differences that can be observed.

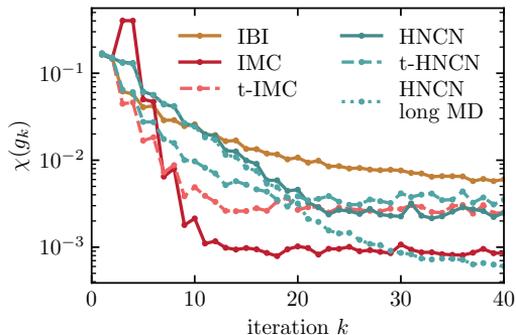


Figure 4.9: Collective convergence of the three RDFs in CG hexane using different coarse-graining methods. At even iterations the non-bonded potentials are updated, uneven iterations only update the bonded potentials. The long MD variant of HNCN uses the same number of MD steps per iteration as IMC ( $2 \times 10^6$  instead of  $2 \times 10^5$ ).

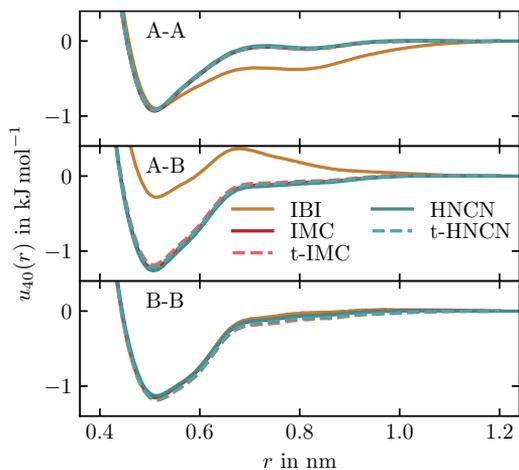


Figure 4.10: The pair potentials  $u_{AA}$ ,  $u_{AB}$ , and  $u_{BB}$  after 40 iterations of IBI, IMC, or HNCN.

---

Next, we tested different Gauss-Newton methods on the hexane system. For the Gauss-Newton scheme, a weighting that puts more weight on the RDF onset and tail is used. As mentioned before,  $r^2/(g_{\text{tgt}} + 10^{-30})$  is used. The numerator ensures a good fit of the RDF in the tail region. In combination with  $r_{\text{res}} > r_{\text{cut}}$  this prevents jumps in the potential at the cut-off which otherwise often occur. The jumps are kept to a minimum by the longer  $r_{\text{res}}$  because they would also lead to a kink in the RDF within the optimized region. The denominator ensures a good fit of the RDF onset region. A small mismatch in this region leads to large fluctuations in the repulsive potential, which can cause instabilities in the MD simulation. The weighting thereby helps to keep the Gauss-Newton iteration stable. The constant  $10^{-30}$  is added to the denominator to ensure the weighting matrix is finite. We note that the software package MagiC[159] also implements the Gauss-Newton method, but without weighting or constraints.

As a first demonstration of the Gauss-Newton method, we optimize the CG hexane potentials with different  $r_{\text{cut}}$  with a constant  $r_{\text{res}}$  of 1.2 nm. The results are shown in fig. 4.11. The potentials with  $r_{\text{cut}} = 0.5$  nm are almost purely repulsive and fail to match the RDFs at larger distances. All other potentials with  $r_{\text{cut}} \geq 0.7$  nm possess pronounced potential wells and match the full RDFs well. The changes in the potentials with the cut-off are not predictable and different from just cutting off the tail of the HNCN potential. This shows the need for the Gauss-Newton type scheme when one is interested in parametrizing short-range potentials. Such models are advantageous for their speed in MD simulations.

We test the three constraints for the Gauss-Newton method introduced in section 4.2.8 on hexane with  $r_{\text{cut}} = 1.2$  nm and  $r_{\text{res}} = 1.6$  nm. The target values are calculated from the atomistic simulation (pressure, intermolecular potential energy) or the atomistic RDF (KBI). The convergence behavior is, in all cases, similarly fast as HNCN, with about 25 iterations to the plateau. In table 4.2 we show the resulting properties of the obtained models. We use  $\chi(u)$  as a gauge of the RDF similarity but note that it is not precisely equivalent to the measure minimized by the Gauss-Newton method with weighting. The results show that all properties are reproduced if they were constrained in the potential update. Also, combinations of constraints succeed in reproducing the target values. This is remarkable since it is not directly possible to combine p-IBI with KB-IBI, which both use a ramp correction.[61] By using the two ramp corrections alternately, it has been possible to achieve a compromise but not a match of the two properties.[62] We can also see the expected negative impact on the match of the RDFs, as the algorithm gives preference to the constraint. The impact varies between constraints and is the smallest for the KBI-constraint and the largest for the combination of the intermolecular potential energy and the pressure constraint. The resulting potentials are shown in fig. 4.12. They show all very similar structures with different potential well depths. The notable exception is *PE-p-t-HNCGN*, which shows a distinctly broader potential well. This deviation from the typical form seems necessary to satisfy both constraints.

Constraining the KBI between different beads of the ABA hexane molecule was done as a demonstration of our method. Physically, it is not very meaningful, since for a longer

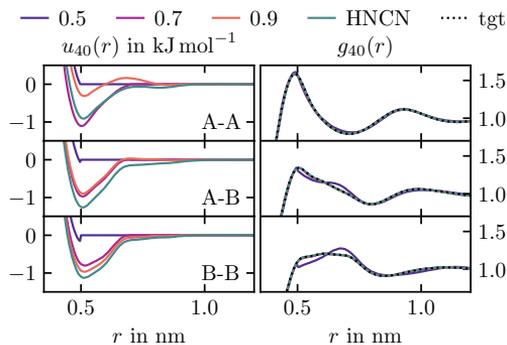


Figure 4.11: Left: A-A, A-B, and B-B potentials in CG hexane optimized for 40 iterations to match the RDFs up to 1.2 nm using HNCN and HNCGN. The numbers in the legend are the values of the cut-off  $r_{\text{cut}}$  in nm used for HNCGN. Right: the resulting RDFs together with the target RDFs.

integration range  $G_{AA}$ ,  $G_{AB}$ , and  $G_{BB}$  are expected to converge to the same value. This was not the case for  $r_{\text{res}} = 1.6$  nm. With mixtures, a KBI constraint can be useful in obtaining models that are more transferable between different concentrations.[62]

We note that so far, we have shown a selection of successful attempts at applying constraints. However, we sometimes found constrained HNCGN and IMC-GN not to converge or fail by generating unphysical potentials. Two examples are: (I) We did not succeed in converging a hexane model with all three constraints. (II) We found it not possible to generate a single-bead water model based on the SPC/E RDF that would match KBI and pressure, similar to the setup of Wang et al., who tried with p-IBI.[61]

Table 4.2: Several properties of coarse-grained hexane models obtained from the Gauss-Newton method with different constraints after 40 iterations. The numbers are averages of the last six iterations and the uncertainties reported are the standard deviation of the six values. The last column gives a measure of the mismatch between the obtained and target RDFs as defined in eq. (4.41)

	$p$ in bar	$PE_{\text{inter}}$ in $\text{kJ mol}^{-1}$	$G_{AA}$ in $\text{\AA}^3$	$G_{AB}$ in $\text{\AA}^3$	$G_{BB}$ in $\text{\AA}^3$	$\chi(g)$ $\times 1000$
target	1.0	-30.4	-207.4	-193.3	-201.9	-
t-HNCGN	1075.5(196)	-12.4(3)	-208.0(1)	-194.0(1)	-202.0(2)	1.0(1)
$p$ -t-HNCGN	2.2(20)	-25.3(1)	-205.7(1)	-191.4(1)	-199.5(2)	2.9(1)
$PE$ -t-HNCGN	-404.1(51)	-30.4(1)	-204.2(1)	-189.8(1)	-197.9(4)	4.2(1)
$KBI$ -t-HNCGN	766.3(288)	-16.0(4)	-207.4(1)	-193.3(1)	-202.1(4)	1.7(3)
$KBI$ - $p$ -t-HNCGN	-0.1(24)	-25.7(1)	-207.3(1)	-193.3(1)	-202.2(3)	3.5(1)
$PE$ - $p$ -t-HNCGN	2.0(21)	-30.4(1)	-207.7(1)	-191.9(1)	-200.1(2)	11.5(1)

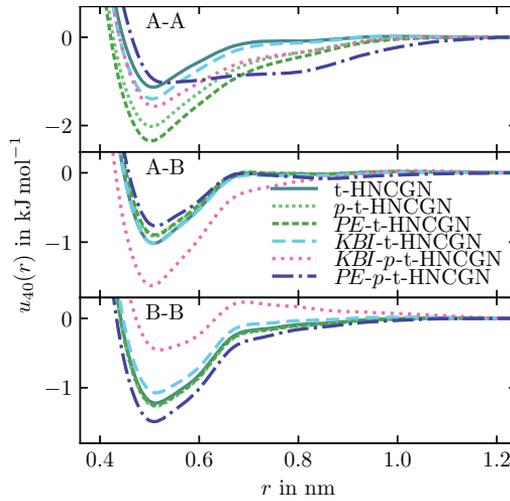


Figure 4.12: The pair potentials for CG hexane after 40 iterations of t-HNCGN with various constraints.

---

---

### 4.4.3 Organic Solvents

In this section, we test our methods on many different systems to learn more about their applicability. We coarse-grain nine organic molecules into single beads. Two different methods are used, t-HNCN and *p*-t-HNCGN, to obtain two models, one which optimally matches the structure and one which has the right pressure and best-possible structure. The convergence behavior is shown in fig. 4.13. We find that for two molecules, pyridine and 1,4-dioxane, the t-HNCN method does not converge but enter some kind of oscillatory behavior between two potentials that both do not reproduce the RDFs. In the plot, this results in prolonged decay of  $\chi(g)$ . For three further molecules, 2-butanol, ethylene glycol, and acetone, the convergence is also significantly slower than for the other molecules. The  $\chi(g)$  value from the initial guess is also the highest for those five molecules. With the *p*-t-HNCGN method, four of those five molecules also show slow or no convergence within 15 steps. The other molecules converge to different  $\chi(g)$  values higher than for t-HNCGN because of the pressure constraint. Note that for the *p*-t-HNCGN method, the convergence is sometimes lower at early iterations, where the pressure is not yet matched.

To further investigate the convergence problems of some molecules we show the direct correlation function in fig. 4.14. The shown functions are the same that were used in the computation of the initial potential guess with eq. (4.9). We see that four of the five molecules with slow convergence also have a local extremum in their direct correlation function at large  $r$ . This is not typical as  $c$  is expected to be short-ranged. The way the direct correlation function is computed (eq. (4.15)), it is assumed to have decayed to zero at half the length of the RDF input. For all molecules, the RDF was calculated up to 3.2 nm. For those molecules, this is not the case and  $c$  has not decayed to zero at 1.6 nm. The local maxima seen in fig. 4.14 are then caused from applying eq. (4.15) on finite RDF data. The intensity of the oscillations in the tail of the RDF, also shown in fig. 4.14, correlates to the occurrence of the local maxima in  $c$ . Ideally, one would calculate longer RDFs and compute the potential guess and  $\frac{\partial c}{\partial h}$  from them. Nevertheless, we can also work with the RDF data available since an approximate Jacobian can still be used to reach convergence. Based on the oscillatory behavior of the potential, we know that the potential update is overshooting. We, therefore, scale the update by 67% and append  $\times 0.67$  to the method name to denote that scaling was used. The results are also shown in fig. 4.13. This leads to convergence of both RDF and pressure in all but one cases: *p*-t-HNCGN $\times 0.67$  for ethylene glycol. For this one molecule, we pre-generate a pair potential by using p-IBI for 15 iterations.[61] Starting from the last p-IBI potential, we find that *p*-t-HNCGN $\times 0.67$  converges.

Furthermore, we parametrize pair potentials for all mixtures of the nine solvents. We decided to keep the pair potentials between alike molecules that were obtained from the coarse-graining of the pure solvents. Therefore, for each mixture, only the mixed interaction potential between the two molecules is optimized as described in section 4.2.7. The target distribution is the RDF of that same interaction up to  $r_{\text{res}} = 1.6$  nm. We use the

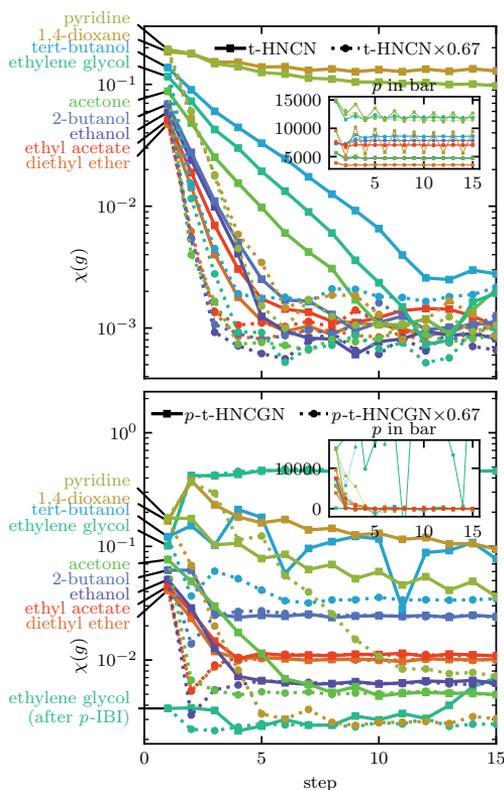


Figure 4.13: The convergence of the RDF for all pure solvents. The insets show the convergence of the pressure with the same color scheme. In the legend,  $\times 0.67$  denotes the scaling factor of the potential update.

t-HNCGN $\times 0.67$  method without constraints, for which the derivative  $\frac{\partial c}{\partial h}$  is determined once from all three target distributions. In section 4.4.3 we show the convergence behavior and the resulting RDF for the mixtures. Different ways to generate an initial guess for  $u_{AB}$  were tested. First we generated an HNC potential guess from eq. (4.9) and only used it for  $u_{AB}$ . This worked for about 80% of the mixtures and would lead to convergence, while the rest would not. For some mixtures, e.g., ethanol + ethylene glycol, the mixture would phase separate at iteration 1 with the initial guess potentials. The following updates were not meaningful, as we assume homogeneity for the integral equation methods. The phase separation of the mixture is caused by the inconsistent combination of different potentials. In other cases, e.g., ethylene glycol + tert-butanol, it would not converge,

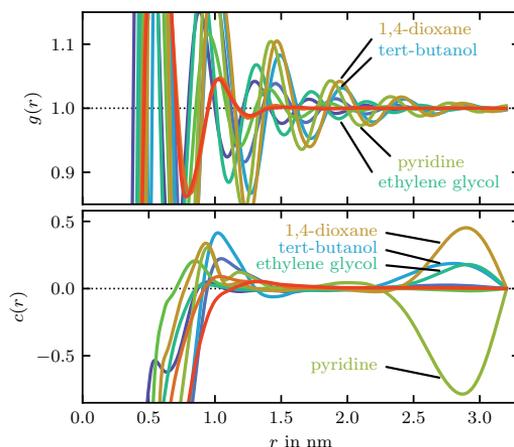


Figure 4.14: The RDFs  $g$  (top) and direct correlation functions  $c$  (bottom) calculated from eq. (4.15) for the nine pure solvents. Four lines have been labeled, the others have the same color as in fig. 4.13. For  $r = 0$  nm,  $c(r)$  reaches values between -60 and -15, and therefore only a sector is shown.

but not due to phase separation of the mixture. Rather, the fluid would partially freeze due to the potential  $u_{AB}$  becoming very attractive. As an alternative initial guess for  $u_{AB}$  we tried  $(u_{AA} + u_{BB})/2$ , denoted as AVG for average in section 4.4.3. This led to similar outcomes as with the HNC initial guess: About 80% of the mixtures would not converge. After those two attempts, seven mixtures that had not clearly converged and required a better initial guess. For those seven systems, we performed 15 pre-iterations of IBI with the HNC initial guess and a potential update scaling of 67%. The final potential was then used and subsequently run for 15 HNC $\times$ 0.67 iterations. We find that IBI often has converged the RDF rather well and the following Newton iterations have almost constant  $\chi(u)$ . The reached convergence for all 36 mixtures shows the wide applicability of our approach, and convergence problems can be attributed to the problem of finding an initial guess to which we have proposed three different methods. We have not performed tests on the transferability of our new potentials but note that with all mixtures converged, one could simulate a system with any composition of the nine molecules.

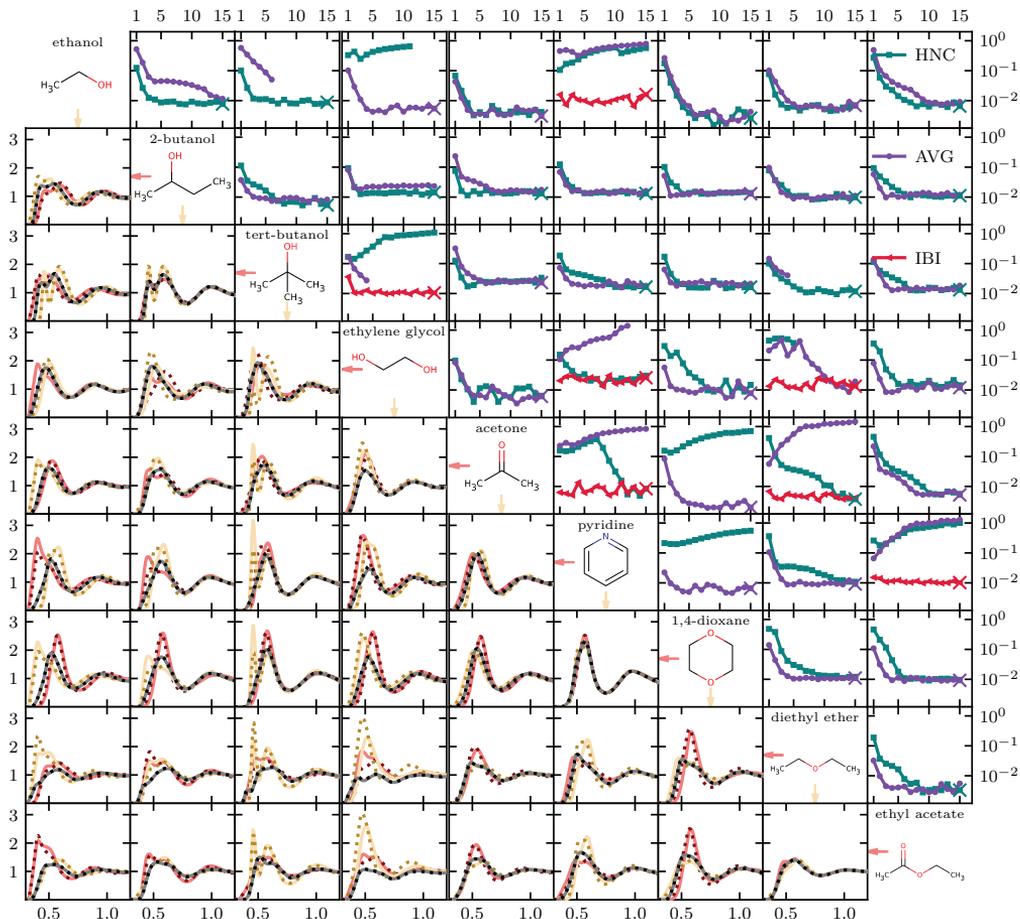


Figure 4.15: On diagonal: The nine organic solvents that were coarse-grained into single beads. Above diagonal: Convergence plots for all mixtures that shows  $\chi(g_{AB})$  plotted against iteration. In all cases, the t-HNCGN $\times 0.67$  method is used to optimize the potential  $u_{AB}$  to match the RDF  $g_{AB}$ , where A and B are the components of the mixture. Different colors and markers indicate different initial guess methods for  $u_{AB}$ . For some mixtures, 15 IBI iterations were performed (not shown), and Newton's method started from the IBI outcome. The x in each plot marks the run that has the best RDF match at iteration 15. Below diagonal: RDFs for all mixtures at iteration 15 of the coarse-graining run that is marked with an x above the diagonal. The distance  $r$  between beads on the x-axis is in nm. The grey, red, and yellow lines are  $g_{AB}$ ,  $g_{AA}$ , and  $g_{BB}$  between the molecule in the same row and column on the diagonal, respectively. Their atomistic target RDFs are shown as dotted lines in black, dark red, and dark yellow.

---

---

## 4.5 Conclusion

We have made progress on three important aspects of structural coarse-graining. Firstly, we have derived and demonstrated how to obtain a fast iterative scheme with integral equation theory for arbitrary molecular systems. It converges similarly fast as IMC for CG systems without bonds and about half as fast if the CG molecule has intramolecular degrees of freedom. By reusing the Jacobian, the cost per iteration is close to that of IBI (which requires far more iterations) and enables the fast derivation of CG models. The reuse of the Jacobian also works for IMC, resulting in an accelerated IMC method. Secondly, we have investigated the instabilities of Newton Methods in the RDF onset region. Due to the exponential relation between RDF and potential in this region, Newton’s methods tend to produce artifacts. We provide a modification for the HNC- and IMC-Jacobian that improves the stability of the methods significantly. This removes the need for regularization or pre-iterations in most cases. Thirdly, the Gauss-Newton formulation of the inverse problem allows for shorter cut-offs in the CG model and thermodynamic constraints to the potential update. We demonstrate how constraints can be derived if they are a function of pair-distance, pair potential, pair-force, and RDF. This is demonstrated for pressure, Kirkwood-Buff integrals, and intermolecular potential energy. The possibility of applying several constraints simultaneously is demonstrated. Finally, we have successfully applied our methods to a large number of organic solvents and their mixtures to demonstrate stability and speed that enable consistent and fast derivation of bottom-up CG force fields.

## Acknowledgements

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within project B1 of the TRR 146 “Multiscale Simulation Methods for Soft Matter Systems” – Project No. 233630050.

## Appendix: Symmetry Adapted Distribution Functions

From the analysis of the trajectory, we obtain the distributions  $h_{\alpha\beta}(r)$  and  $e_{\alpha\beta}(r)$  already average over the indistinguishable beads of type  $\alpha$  and  $\beta$ , but not yet symmetry-adapted. The distribution function  $e(r)$  is the intramolecular RDF. Both  $h(r)$  and  $e(r)$  can be obtained with our extension to VOTCA’s `csg_stat` tool. Here we derive the equations needed to obtain  $H_{\alpha\beta}$  from  $h_{\alpha\beta}$  and  $\Omega_{\alpha\beta}$  from  $e_{\alpha\beta}$ .

In section 4.2.5 we introduced the matrix  $\mathbf{h}'$  which has individual rows and columns for each bead. In contrast,  $\mathbf{h}$  has one RDF per bead type combination. The elements in each block of  $\mathbf{h}'$  are equivalent to a single element in  $\mathbf{h}$

$$(\mathbf{h}'^{\alpha\beta})_{kl} = h_{\alpha\beta} \quad \text{independent of } k, l. \quad (4.42)$$

Here,  $h_{\alpha\beta}$  is an element of  $\mathbf{h}$  and  $\mathbf{h}'^{\alpha\beta}$  is a block of  $\mathbf{h}'$ .  $\mathbf{H}$  and  $\mathbf{h}$  are already of the same dimensionality, but the weighting that comes from the transformation matrix  $\mathbf{T}$  still has to be applied. The first column of the transformation matrix (see section 4.2.4) has its elements given by

$$T_{k1}^{\alpha} = n_{\alpha}^{-\frac{1}{2}} \quad \text{for any row } k. \quad (4.43)$$

If we apply the transformation, insert eq. (4.42), and calculate the top left element we obtain

$$\begin{aligned} H_{\alpha\beta} &= (\mathbf{H}^{\alpha\beta})_{11} = ((\mathbf{T}^{\alpha})^{-1} \mathbf{h}'^{\alpha\beta} \mathbf{T}^{\beta})_{11} \\ &= \left( \sum_{k=1}^{n_{\alpha}} n_{\alpha}^{-\frac{1}{2}} \right) \left( \sum_{l=1}^{n_{\beta}} n_{\beta}^{-\frac{1}{2}} \right) h_{\alpha\beta} \\ &= (n_{\alpha} n_{\beta})^{\frac{1}{2}} h_{\alpha\beta}. \end{aligned} \quad (4.44)$$

Here we have used the property  $(\mathbf{T}^{\alpha})^{-1} = (\mathbf{T}^{\alpha})^{\text{T}}$  of orthonormal matrices. The same prefactor of  $(n_{\alpha} n_{\beta})^{\frac{1}{2}}$  applies for the relation between the elements in  $\mathbf{C}$  and  $\mathbf{c}$ .

For the intramolecular density function, the starting point is the definition of  $\omega'_{kl}$  between beads  $k$  and  $l$ . It is most easily defined in Fourier space

$$\hat{\omega}'_{kl} = \delta_{kl} + \rho_l \hat{e}'_{kl}. \quad (4.45)$$

Here,  $e'_{kl}(r)$  is the intramolecular RDF between bead  $k$  and  $l$ . The density  $\rho_l$  is the number density of bead  $l$ . The first term in eq. (4.45) represents the delta-peak of the reference atom, which is included in  $\omega'$  but not in  $e'$ . The elements of  $\omega$  are the average over a row in a block from  $\omega'$  (see Ref. [149])

$$\omega_{\alpha\beta} = \frac{1}{n_{\beta}} \sum_{l=1}^{n_{\beta}} (\omega'^{\alpha\beta})_{kl} \quad \text{for any row } k. \quad (4.46)$$

The same relation holds between  $\mathbf{e}'$  and  $\mathbf{e}$ . Combining eqs. (4.45) and (4.46) and using  $n_{\beta} \rho_l = \rho_{\beta}$ , we obtain

$$\hat{\omega}_{\alpha\beta} = \frac{1}{n_{\beta}} (\delta_{\alpha\beta} + \rho_{\beta} \hat{e}_{\alpha\beta}). \quad (4.47)$$

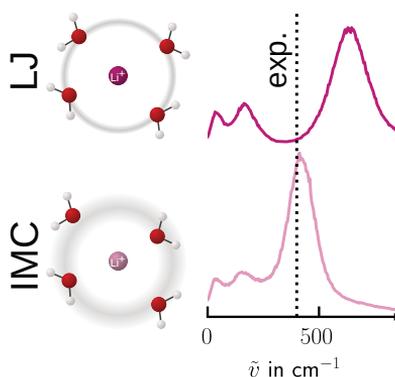
Applying the transformation matrix to a single block of  $\omega'$  gives

$$\begin{aligned} \hat{\Omega}_{\alpha\beta} &= (\hat{\Omega}^{\alpha\beta})_{11} = ((\mathbf{T}^{\alpha})^{-1} \hat{\omega}'^{\alpha\beta} \mathbf{T}^{\beta})_{11} \\ &= \left( \sum_{k=1}^{n_{\alpha}} n_{\alpha}^{-\frac{1}{2}} \right) \left( \sum_{l=1}^{n_{\beta}} n_{\beta}^{-\frac{1}{2}} \right) \hat{\omega}_{\alpha\beta} \\ &= (n_{\alpha} n_{\beta})^{\frac{1}{2}} \hat{\omega}_{\alpha\beta} \\ &= (n_{\alpha}/n_{\beta})^{\frac{1}{2}} (\delta_{\alpha\beta} + \rho_{\beta} \hat{e}_{\alpha\beta}). \end{aligned} \quad (4.48)$$

---

## 5 Where Lennard-Jones Potentials Fail: Iterative Optimization of Ion-Water Pair Potentials Based on Ab Initio Molecular Dynamics Data

---



### Abstract

The use of the Lennard-Jones (LJ) potential in computer simulations of aqueous electrolyte solutions is widespread. The standard approach is to parameterize LJ potential parameters against thermodynamic solution properties but problems in representing the local structural and dynamic properties of ion hydration shells remain. The  $r^{-12}$ -term in the LJ potential is responsible for this as it leads to overly repulsive ion-water interactions at short range. As a result, the LJ potential predicts blue shifted vibrational peaks of the cations' rattling mode and too large negative ion hydration entropies. We demonstrate that cation-water effective pair potentials derived from ab-initio MD data have softer short-range repulsions and represent hydration shell properties significantly better. Our findings indicate that replacing the LJ potential with these effective pair potentials offers

---

---

a promising route to represent thermodynamic solution properties and local interactions of specific ions with non-polarizable force fields models.

## 5.1 Letter

The properties of aqueous electrolyte solutions are well understood qualitatively but remain difficult to model quantitatively.[160] The Lennard-Jones (LJ) 12-6 potential, combined with electrostatic interactions between non-polarizable charges, is a popular choice for the intermolecular ion-water potential used in classical molecular dynamics (MD) simulations. Its parametrization can target a large list of experimentally available properties.[161] The target properties include ion solvation free energy and entropy, solubility, radial distribution function (RDF) peak positions, osmotic pressure, solution density, and also transport properties such as viscosity and diffusion coefficient.[21, 22, 162] However, ion solvation entropies have often been considered secondary and spectroscopic properties are rarely compared with experiments. Representability problems, such as very low solubility, have been ascribed to electronic polarization effects caused by the high charge density of the ions.[163, 164] Polarization can be introduced explicitly into molecular models.[165, 166] As an alternative to introducing polarizability explicitly, a mean-field electronic continuum correction (ECC) has been proposed, based on scaling of the ionic charges.[17, 167–169] Furthermore, many ion force fields break the LJ combining (mixing) rules and provide separate parameters for ion-water and ion-ion interactions.[22, 23, 170–173] Other groups reported success for multivalent ions by extending the LJ potential with an  $r^{-4}$  term.[174]

While the use of charge scaling and other approaches to account for the lacking electronic polarizability of the models is a topic of debate,[175, 176] the  $r^{-12}$  form of the short-range repulsion remains mostly untouched.

This is surprising, since already two decades ago quantum-mechanical calculations revealed that the ion-oxygen interaction is much better described by an exponential or  $r^{-7}$  term.[177, 178] Recently, it was demonstrated that experimentally measured dynamical and structural properties of ion hydration shells are difficult to reproduce with ion-water LJ potentials.[179] While commonly used in computer simulations of solids and molten salts, an exponential repulsion term is rarely used in computer simulations of electrolyte solutions.[180–182] For molten salts in different phases, several properties are better reproduced by a Buckingham than by a LJ potential.[183]

In this Letter, we answer the question of how good a choice the LJ potential is for describing the ion-water interaction. More specifically, we aim at answering the question of which properties of electrolyte solutions can or can not be reproduced with existing models. One can expect the steepness of the repulsive short-ranged potential to have a direct influence on the dynamics of the ion and its solvation shell. Those dynamics are highly relevant, since there is ongoing progress in measuring and interpreting terahertz

---

---

spectra of ion solutions.[184–186]

To find a free-form description of the ion-water interaction we followed a methodology that has been first described by Lyubartsev.[25] From ab-initio MD (AIMD) simulations of a single ion solvated by water molecules, we obtained the ion-oxygen RDF. Using the iterative Boltzmann inversion (IBI) and inverse Monte-Carlo (IMC) methods, we subsequently derived effective pair potentials that reproduce the RDF in classical MD simulations.[50, 52, 187] The simulation details are provided in the SI together with the IBI-IMC procedures (sections S1-S3). Lyubartsev found that the resulting lithium-oxygen interaction can best be described by an exponential potential and is much softer than a LJ description. In this study, we investigate LiCl, NaCl, KCl, and CaCl<sub>2</sub>, and thus four different cations representing strongly and weakly hydrated cations in the cationic Hofmeister series. For comparison and as basis for our improved force fields we considered four different LJ parameter sets: OPLS,[188, 189] ECC,[19, 190, 191] HMN,[21, 172] and Madrid.[22] The OPLS, ECC, and HMN ion parameters were used together with the SPC/E water model,[16] the Madrid parameters were used in combination with the TIP4P/2005 water model.[192] We adapted the ion-oxygen potentials of ECC and HMN in a tabulated form with iterative methods to reproduce the AIMD RDFs. The resulting force fields are referred to as ECC IMC and HMN IMC. To estimate the “repulsive steepness” of the obtained pair-potentials we considered two possible routes. In the first route, the potential was fitted in a single step with a function of form  $u_{\text{fit}}(r) = \frac{C_{\text{rep}}}{r^n} - \frac{C_6}{r^6}$ , providing the exponent  $n$ . In the second route, the attractive part of the potential was fitted in a first step and the remaining repulsive part in a second step (SI, section S4). We found similar results for both routes, but only show the results of the second because it allows for a clearer presentation. Figure 5.1 shows  $u_{\text{rep}}(r)$  and  $u_{\text{rep,fit}}(r)$  in a double-logarithmic plot. We find exponents that lie between 6.9 for the HMN IMC model of the O-Li interaction up to 13.2 for O-Cl in the ECC IMC model. For the full charge HMN IMC model, all exponents for cation-oxygen interactions are between 6.9 and 9.8, indicating a much softer repulsion as in a typical LJ model. For the ECC IMC model, the exponents are more diverse, ranging from 8.9 for O-Li up to values larger than 12 for O-K and O-Cl. A general trend seems to be the decrease of the exponent with increasing charge density of the cation.

We now address the question why the form of the short-range ion-water repulsion is especially important in electrolyte systems. In apolar systems at ambient pressure, the RDF has a first peak that corresponds to the position of the minimum of the LJ potential at  $2^{1/6}\sigma$ . [193] The situation changes when two atoms have opposite charges and the two atoms are attracted by the Coulomb potential. This moves the first peak of the RDF far into the repulsive region of the LJ potential and makes the repulsive part of the LJ potential more important.[161] This is illustrated in the SI, section S5. At the same time, the attractive part of the LJ potential has negligible influence, since the electrostatic attraction is much larger. This results in an interdependence of  $\epsilon$  and  $\sigma$  in the repulsive part which complicates the parametrization of force fields. In the parameter space, lines

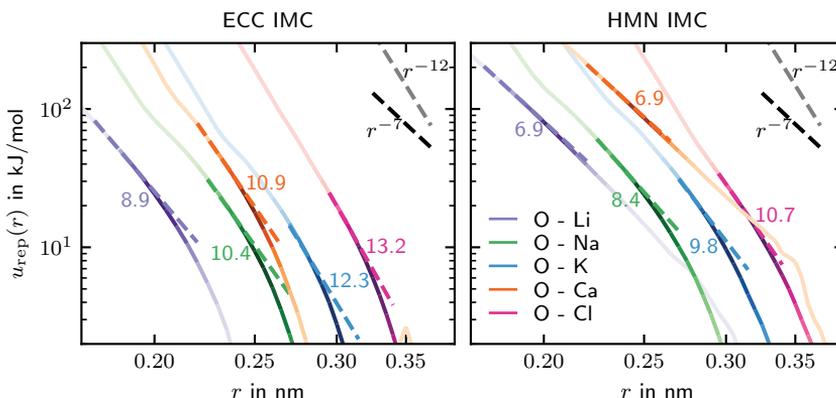


Figure 5.1: Double-logarithmic plot of the repulsive part of the obtained ion-oxygen potential. The potentials are depicted as solid lines, shaded darker for higher values of the associated RDF. Exponents  $n$  of the power function fits of form  $\frac{C_{\text{rep}}}{r^n}$  are shown in the respective color next to the curve. The fits are depicted as short dashed lines.

with  $\epsilon\sigma^{12} = \text{const.}$  can be expected produce similar outcomes. We see this for example in the work of Horinek et al. [21] (figures 16 and 17) or Mamatkulov et al. [172] (figures 4 and 6).

Our results up to this point indicate three things. First, the two parameters of the cation-oxygen LJ potential are interdependent and different  $(\sigma, \epsilon)$  combinations provide models with comparable accuracy. Second, the repulsive  $r^{-12}$  form is unsuited to reproduce the first peak of the ion-water RDF from AIMD data (fig. 5.3 and SI, figure S8). Classical force field parametrization of ions therefore yields different LJ parameter sets, dependent on what experimental properties are fitted. This is supported by the observation that ion LJ parameters vary widely whereas parameters for atoms with zero or nonzero partial charges are similar across different force fields. Third, we expect that our tabulated pair potentials reproduce dynamic and thermodynamic ion hydration shell properties in better agreement with experimental data than LJ potentials.

We first focus on the properties that are not generally considered in the context of force field optimization of ions: spectroscopically measurable vibrational dynamics. Ions and the water molecules in the solvation shell perform several active vibrations, which lie in the terahertz regime. For  $\text{MgCl}_2(\text{aq})$  and  $\text{CaCl}_2(\text{aq})$ , the highest frequency peak introduced by the cation has been identified as a normal mode in which the cation and one of the solvation shell oxygen atoms oscillate against each other.[185] To compare with experimental data, we computed the vibrational density of states (DOS) of the ions in the

salt solution (SI, section S6). We obtained spectra containing two or three distinct peaks (figure S6) and fitted them with the sum of two or three Lorentzian functions. In fig. 5.2a, we compare selected frequencies to experimental data for the LiCl, NaCl, and CaCl<sub>2</sub> systems. See fig. 5.3 for the cation spectra of LiCl and CaCl<sub>2</sub> solutions and a selection of force fields. Further spectra are shown in figure S6. To confirm that the rightmost peak in the frequency range up to 800 cm<sup>-1</sup> shown in figures 5.2a and S6 always corresponds to the ion rattling against the solvation shell, we also computed the cation-oxygen potential of mean force and used the harmonic approximation to its first minimum (SI: section S6, figure S7). The corresponding frequencies are shown in fig. 5.2a.

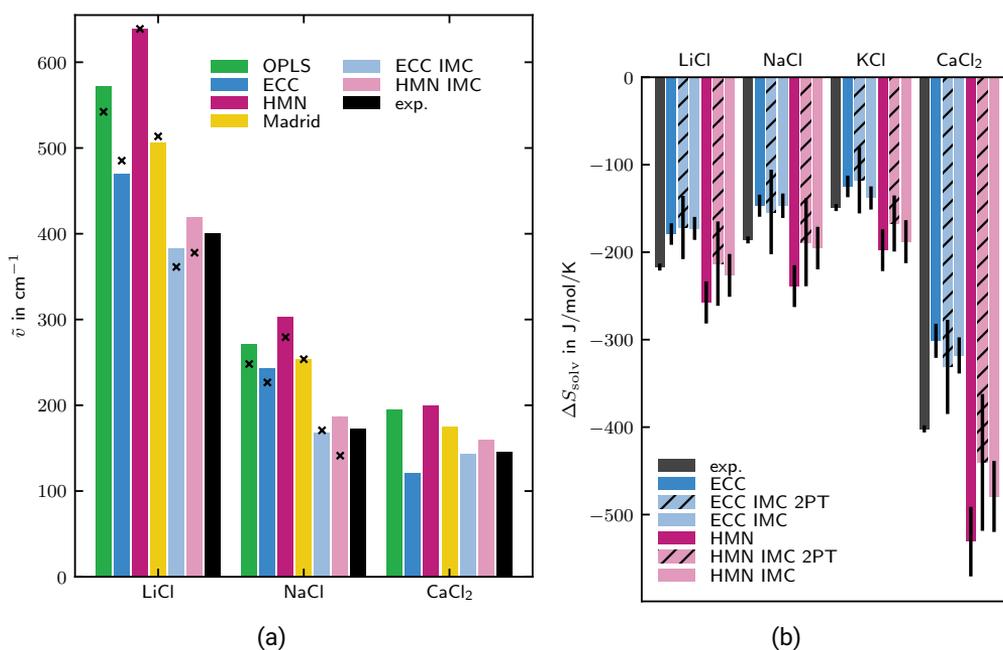


Figure 5.2: (a) The peak frequency of the rattling motion of cations in their solvation shell. This is the highest frequency peak for LiCl and NaCl and the second highest for CaCl<sub>2</sub>. For LiCl and NaCl the frequencies obtained from the harmonic approximation of the first minimum of the potential of mean force between the cation and oxygen are shown as black crosses. The experimental data is taken from Schwaab et al.[186] Similar experimental data is reported by Schmidt et al.[184] (b) The solvation entropies obtained with the ECC and HMN force fields in comparison to experimental data. Error bars for ECC IMC and HMN IMC indicate the standard deviation of five independent calculations.

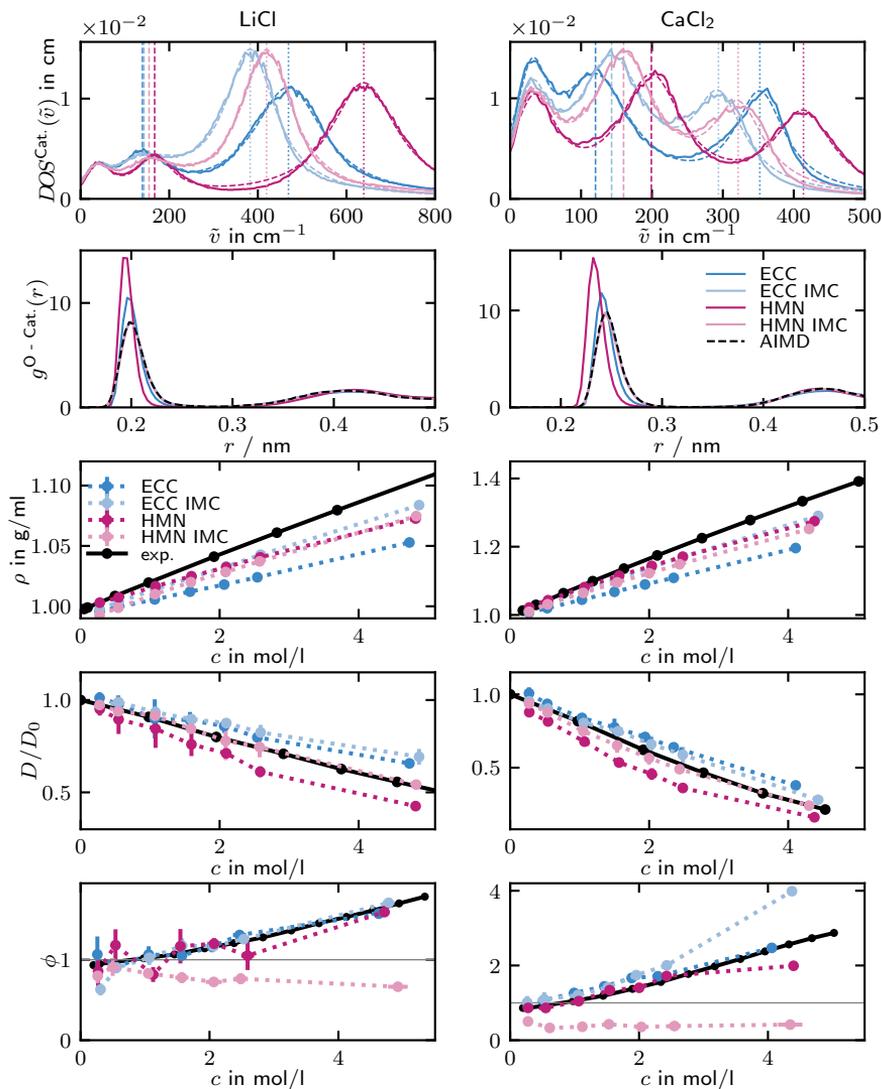


Figure 5.3: Spectroscopic and thermodynamic solution properties. From top to bottom: cation DOS, cation-water RDF, mass density  $\rho$ , relative water diffusivity  $D/D_0$  ( $D_0$  is the self-diffusion coefficient of water at zero salt concentration), and osmotic coefficient  $\phi$  for LiCl (left panels) and  $\text{CaCl}_2$  (right panels). In the DOS plots, the dashed lines represent the fits by a sum of three Lorentzian distributions and the vertical dotted and dashed lines mark the highest and second highest frequency peak, respectively.

---

---

We found LJ force fields to consistently overestimate the frequency of the rightmost peak for LiCl and NaCl (figure 3 and figure S6). On the other hand, the IMC improved force fields reproduced those frequencies very accurately. Considering our earlier findings above, this can be interpreted as the LJ interaction having a too steep repulsive term, which induces too stiff normal modes. Among the LJ models, the ECC model performed the best but is still systematically off. For  $\text{Ca}^{2+}$ , where we compare the second highest frequency, the ECC model prediction is redshifted with respect to the experiment. Note that in this case a redshifted peak is corrected in the appropriate direction, while with HMN this is achieved for a blueshifted one, indicating how important the correct representation of the first solvation shell is for the dynamics in electrolytes.

Motivated by these results we turn to a thermodynamic property that should be influenced directly by overly steep repulsion: the ion solvation entropy. In their systematic scanning of the LJ parameters for ions, Horinek et al. found that there is no combination of  $\sigma$  and  $\epsilon$  that can reproduce the ion solvation free energy and the solvation structure on one side, and solvation entropy on the other side.[21] Specifically, they found that the solvation entropies of several ion models described with LJ parameters are systematically too negative. Consequently, many force fields share this trend.[194] The stiffness of the ion-water interaction lies at the origin of this offset. Figure 5.2a indicates that the LJ models predict higher peak frequencies than the IMC models among the pairs examined here. This trend can be related to changes in ion hydration entropy, i.e. a blue shift in peak frequency correlates with a decrease of the entropy. One model that enables this connection semi-quantitatively is the 2PT model.[195, 196] We used it to calculate the solvation entropy ( $\Delta S_{\text{solv}}$ ) of the ECC IMC and HMN IMC model (calculation details can be found in the SI, section S7). The results are shown in fig. 5.2b. The data indicate that the solvation entropy obtained with the HMN IMC model is indeed less negative than the solvation entropy obtained with the HMN model. We further carried out thermodynamic integration (TI) to obtain more precise values of the solvation entropy (more details in the SI, section S8). The results, also shown in fig. 5.2b, demonstrate that with the HMN IMC potentials the experimental solvation entropies are matched better. This is remarkable because Horinek et al. showed that there is no combination of Lennard-Jones parameters that can match the solvation free energy (see SI, figure S8) and solvation entropy for halide ions. This indicates that LJ 12-6 potentials might be too steep to faithfully represent the solvation entropy of several salts. For the ECC and ECC IMC model, a moderate improvement is achieved for KCl and  $\text{CaCl}_2$  but overall the solvation entropies are still too small in absolute magnitude. This indicates that the charge scaling underestimates the ion-water attraction. Due to the water molecules being bound less tightly, the solvation entropies are less negative than the HMN model. Improvement of the short-range ion-water repulsion achieved by the ECC IMC model has, in this case, a smaller effect on the ion hydration entropies. Note that this correlates with the observation that the exponents  $n$  of  $u_{\text{rep,fit}}(r)$  are generally closer to 12 for ECC IMC than for HMN IMC (fig. 5.1).

To validate our derived potentials on more common properties we calculated the ion-oxygen RDF, density ( $\rho$ ), water self-diffusion coefficients ( $D$ ), and osmotic coefficients ( $\phi$ ). The latter three quantities were calculated for different salt concentrations and compared directly with experiment. For those four properties, we each calculated a root mean square distance (RMSD),  $\Delta_x$ , to the experimental values, averaged over all concentrations and salts. For the RDF,  $\Delta_{\text{RDF}}$  takes the AIMD result as reference and was only computed for the lowest concentration. Those values can be found in table 5.1. We show the data for LiCl and CaCl<sub>2</sub> and a selection of force-fields in fig. 5.3 (data for NaCl and KCl and all force fields is shown in the SI, section S9). It can be observed that the

Table 5.1: RMSD measures for the reproduction of reference values from experiment ( $\rho$ ,  $D/D_0$ ,  $\phi$ ) or AIMD (RDF).

	$\Delta_{\text{RDF}}$	$\Delta_\rho$ (g mL <sup>-1</sup> )	$\Delta_{D/D_0}$	$\Delta_\phi$
OPLS	1.38	0.046	0.45	1.43
ECC	1.07	0.166	0.28	0.61
HMN	2.29	0.050	0.63	0.54
Madrid	2.17	0.007	0.51	0.57
ECC IMC	0.06	0.071	0.24	0.77
HMN IMC	0.12	0.096	0.30	2.87

RDF is much better reproduced in the IMC improved force fields. This is to no surprise as that is the objective of the iterative procedure. From the LJ force fields, HMN and Madrid depart furthest from the AIMD solvation structure.

The experimental density trends with concentration are best reproduced by the Madrid force field. This reflects that the density of the salt solution was an essential part of the Madrid force field optimization strategy.[22] We observed that ECC IMC is an improvement in comparison with the ECC model. At the same time, the HMN IMC model performs worse at reproducing the density in comparison with the HMN model.

For the water self-diffusion, it is evident that the ECC models with scaled charges perform best. When taking a detailed look at the data as presented in the SI, figure S11, it becomes clear that only ECC models correctly predict a slight acceleration of water diffusion in dilute KCl solutions. Earlier papers assumed this qualitative trend could only be modeled with different water models or explicit electronic degrees of freedom would be necessary.[197, 198] Since we find no significant change between ECC and ECC IMC the charge scaling seems to be the important factor, aligned with earlier findings.[199] However, the Madrid model, which has scaled charges (0.85) and uses the TIP4P/2005 water model instead of the SPC/E water model, does not reproduce the qualitative trend. So the choice of the water model has non negligible impact. In addition to this clear advantage of the ECC models, the IMC improved force fields score better than their base

---

---

LJ counterparts.

The experimental osmotic coefficients are similarly well reproduced by all force fields except by OPLS and HMN IMC. We observed that the HMN IMC model predicts too strong ion-pairing which led an underestimation of the osmotic pressure. The optimization of only the ion-water interactions apparently changed the balance of ion-water and ion-ion interactions in the HMN force field. To achieve better agreement, a re-parameterization of the HMN ion-ion interactions with a Kirkwood-Buff or related approach may therefore be considered. [23, 191] We note that, alternatively, multistate reference ensembles[129] may be used in a IBI/IMC approach that combines the ion-water RDF from AIMD calculations and ion-ion RDFs from Kirkwood-Buff-derived nonpolarizable force fields.

In summary, we have found that the LJ 12-6 potential has several problems when used for modeling ion-water interactions with nonpolarizable force fields. Its parameters are highly correlated leading to their reported values being widely scattered; even when optimizing for the same experimental properties. Significantly, the LJ potential has a non-correctable bias that precludes the correct representation of the ion hydration entropy and its vibrational dynamics in the terahertz frequency range. Based on the iterative optimization method used herein, we found that the short-range decay of the ion-water effective pair potential does not follow a  $1/r^{12}$  dependency, but decays slower and is better represented by a Buckingham potential or a  $m$ -6 Mie potential (with  $m < 12$ ), in particular for cations with a high charge density. We expect this insight to bring forward the application of classical molecular dynamics simulations in the field of Hofmeister ion chemistry. The approach used herein can be extended to deriving effective pair potentials for interactions that cannot readily be described with standard combination rules.

## Acknowledgement

We wish to thank Azade Yazdanjar for fruitful discussions. We are grateful for financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB TRR 146 (Grant 233630050, project B1) and the MaxWater initiative of the Max Planck Society.

## 5.2 Supporting Information

Details on ab initio molecular dynamics, classical molecular dynamics, inverse methods, potential fitting, LJ potentials and high Coulomb attraction, ion vibrational density of states, 2PT model entropy, and thermodynamic integration, Full data for RDF, density, water diffusion, and osmotic coefficient.

---

## 5.2.1 Ab Initio Molecular Dynamics

All ab initio MD simulations have been carried out with CP2K package.[200] The systems contained 100 water molecules and 1 ion pair for monovalent salts (LiCl, NaCl, KCl), and 160 water molecules and 1 ion pair for the divalent salt (CaCl<sub>2</sub>). The cubic box size was determined by fixing the density of ion solution to 1.0 g cm<sup>-3</sup>. For ab initio calculation, we used the revPBE exchange-correlation functional[201] together with Grimmes' van der Waals corrections.[202]

We have used the mixed Gaussian and plane wave approach as implemented in the CP2K code. The TZV2P basis set which is constructed using triple-valence  $\zeta$  Gaussian basis with two sets of polarization functions was used. We set the plane wave density cutoff of 320 Ry. Norm-conserving Goedecker-Teter-Hutter pseudopotentials[203] were used to describe the core electrons. We set the time step for integrating equation of motion to 0.5 fs. All simulations were performed at 300 K in NVT ensemble with the thermostat of the canonical sampling through velocity rescaling method.[153] The trajectory was written every 5 femto seconds.

We equilibrated the configuration for 15 ps to 20 ps and subsequently we generated 10 ps to 12 ps trajectories, which were used for analysis. The resulting RDF are shown in fig. 5.4. To check for proper equilibration we performed seven independent runs with 20 ps to 22.5 ps length. Their starting position and velocities were taken from the equilibrated trajectory. The resulting RDFs are shown in fig. 5.4 and they confirm that our generated RDFs are reasonably converged.

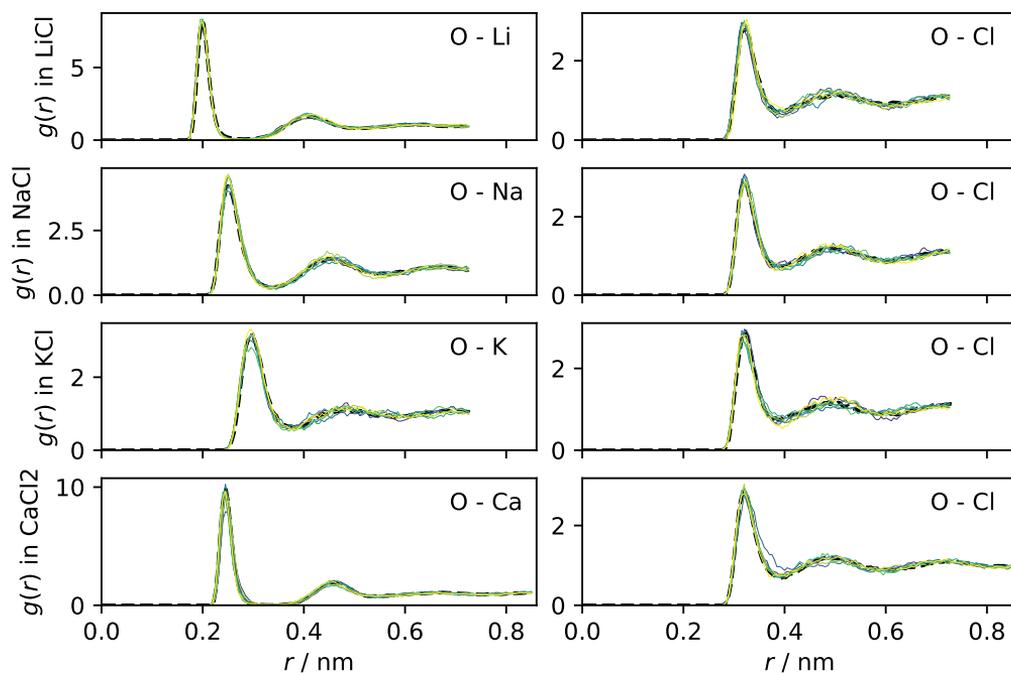


Figure 5.4: The calculated RDF from AIMD simulations. The dashed line is the result from the production run. The colored solid lines are the results from the seven independent additional runs.

## 5.2.2 Classical Molecular Dynamics

All classical MD simulations were carried out with the GROMACS 2019 package.[139] For the equilibrium simulations at constant pressure, we started with a random configuration of water and ions in a cubic box, performed energy minimization, and equilibrated in several steps over a total of 0.4 ns. Production runs of 0.8 ns were performed using the Parrinello-Rahman barostat (1 bar, time constant 5 ps) and a stochastic integrator for constant temperature (300 K, time constant 2 ps). Neighbor searching was handled by a bufferend pair list with an estimated energy drift of  $0.005 \text{ kJ mol}^{-1} \text{ ps}^{-1}$  per particle.[204] MD simulation were performed for four salts LiCl, NaCl, KCl,  $\text{CaCl}_2$  and for 25, 50, 100, 150, 200, 250, 500 formula units of salt in 5000 water molecules. Force field details, including cut-offs, dispersion correction, and water model are given in the following list and tables 5.2 and 5.3. A time step of 2 fs was used for all simulations. Electrostatics were treated with the Particle-Mesh Ewald method with the same real-space cut-off as

---

---

used for the short-ranged interactions and 48 grid points per dimension.

OPLS The ion parameters are based on the work of Chandrasekhar[188] and Åqvist [189], adjusted to the geometric combining rules.

ECC  $\text{Li}^+$ ,  $\text{Na}^+$ ,  $\text{Cl}^-$  parameters were taken from from Kohagen et al.[190],  $\text{Ca}^{2+}$  from Martinek et al.[191], and  $\text{K}^+$  from Bruce et al.[19]. All ion charges were scaled by 0.75.

HMN All but  $\text{Ca}^{2+}$  parameters were taken from Horinek et al.[21] (parameter set 5b with  $\epsilon = 0.65$  kJ/mol).The  $\text{Ca}^{2+}$  parameters were taken from Mamatkulov et al.[172].

Madrid This force field has ion charges scaled by a factor 0.85 and breaks combining rules for several ion-ion and ion-atom combinations.[22]

The ECC IMC and HMN IMC force fields use all parameters of their base LJ model and only have modified, tabulated potentials for the ion-oxygen interactions. For the IMC force fields with tabulated potentials a simple pair list based on atom groups was used and updated every 10 steps. The IMC models do not use a tail correction.

Table 5.2: Details on the force fields used in this paper. For the Madrid force field, there are specific LJ parameter for each interaction, given in table 5.3. The column tail-corr. indicates whether a dispersion correction is applied to the energy and pressure of the system.

	cut-off in nm	tail corr.	water model	mixing rule		$\sigma$ in nm	$\epsilon$ in $\text{kJ mol}^{-1}$	$q$ in $e$
OPLS	0.9	yes	SPC/E	geom.	Ca	0.241203	1.881360	2.00
					K	0.493463	0.001372	1.00
					Li	0.212645	0.076479	1.00
					Na	0.333045	0.011598	1.00
					Cl	0.441724	0.492833	-1.00
ECC	1.2	no	SPC/E	Lor.-Ber.	Ca	0.266560	0.507200	1.50
					Li	0.180000	0.076470	0.75
					K	0.334000	0.130000	0.75
					Na	0.211500	0.544284	0.75
					Cl	0.410000	0.492800	-0.75
HMN	0.9	yes	SPC/E	Lor.-Ber.	Ca	0.241000	0.940000	2.00
					K	0.289000	0.650000	1.00
					Li	0.147000	0.650000	1.00
					Na	0.223000	0.650000	1.00
					Cl	0.440000	0.420000	-1.00
Madrid	1.0	yes	TIP4P /2005	Lor.-Ber.	Ca	0.266560	0.507200	1.70
					K	0.230140	1.985740	0.85
					Li	0.143970	0.435090	0.85
					Na	0.221737	1.472356	0.85
					Cl	0.469906	0.076923	-0.85

Table 5.3: Non-combining-rule values for the Lennard-Jones potentials between atoms and ions in the Madrid model.

	$\sigma$ in nm	$\epsilon$ in kJ mol <sup>-1</sup>
O - Cl	0.42386698	0.06198347
Na - O	0.26083754	0.79338830
Na - Cl	0.30051231	1.43889423
O - Ca	0.24000000	7.25000000
Ca - Cl	0.31500000	1.00000000
O - K	0.28904000	1.40043000
K - Cl	0.33970000	1.40000000
O - Li	0.21200000	0.70065003
Li - Cl	0.27000000	1.28294385

### 5.2.3 Inverse Methods

We employed the software package VOTCA to perform the IBI and IMC iterations.[87] For the iterative procedure, the box size was chosen to match the average from the respective 1 bar NPT runs of the unmodified model. The volume was constant (NVT) during the iterations. We did not employ a pressure correction algorithm.[61] This choice leads to pressures higher or lower than 1 bar and, when used in NPT simulations, to small changes in the density.

The iterative potential improvement was performed in a seven-step procedure: steps 1, 3, 5, 7 are IBI iterations and 2, 4, 6 are IMC iterations. In all steps, 5 iterations of the respective potential update were performed except steps 6 and 7 which each performed 10 iterations. This scheme was found to be useful as an ad hoc solution to several issues when using only IBI or IMC for the task. (i) Pure IBI is sometimes very slow in converging potentials. (ii) IMC can be unstable when the starting point is far from the target (iii) IMC can have severe instabilities when the current or target RDF have differently large regions with a value of zero. For the CaCl<sub>2</sub> system starting from the HMN force field an extra step with 10 IMC iterations had to be performed in order to reach satisfactory convergence.

We took the pair-potential from the last IMC iterations of the ECC systems. For the HMN systems, the fluctuations were much larger and the iteration with the lowest deviation from the AIMD RDF was chosen. This is in itself an interesting observation, as it hints that for ECC systems with scaled charges the space of all possible potentials is more steeply pointing towards a single optimal interaction form. Finally, the oxygen-chloride potentials from the four salts were averaged arithmetically.

In fig. 5.5 we show the convergence of the measure  $\Delta^{\text{cat.-O}} + \Delta^{\text{an.-O}}$ , where  $\Delta$  is the root mean square difference of the current and target RDF

$$\Delta = \sqrt{\frac{1}{r_{\text{max}}} \int_0^{r_{\text{max}}} (g_{\text{cur}}(r) - g_{\text{tgt}}(r))^2 dr}. \quad (5.1)$$

Here,  $g_{\text{cur}}(r)$  and  $g_{\text{tgt}}(r)$  are the current and target RDF, respectively and  $r_{\text{max}}$  is the maximum radius available from the AIMD data. The latter is 0.726 nm for LiCl and NaCl, 0.73 nm for KCl, and 0.85 nm for CaCl<sub>2</sub>.

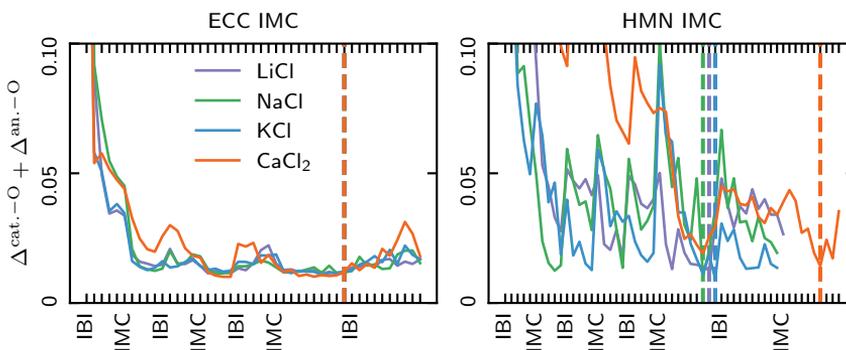


Figure 5.5: The root mean square distance  $\Delta$  of the target AIMD RDF and the RDF of each iteration. The value presented is the sum of the anion-oxygen and cation-oxygen interaction for four different salt solutions. Vertical dashed lines denote the iteration of which the potentials have been used for further analysis.

The tabulated potentials and the AIMD RDF are made available at <https://github.com/marvinbernhardt/ion-water-potentials>.

## 5.2.4 Potential Fitting

The effective pair potentials obtained with IMC were fitted in two steps. Note that this was done only to obtain information on the steepness of the repulsion; MD simulations were performed with the unfitted IMC potentials. A two-step fitting process is possible because the repulsive part of the potential decays quickly and has little influence on the attractive tail and vice versa. As a first step the attractive part of the potential  $u_{\text{att}}(r)$  was fitted from the minimum to the cut-off  $r_{\text{max}}$  with  $u_{\text{att,fit}}(r) = -\frac{C_6}{r^6}$ . The repulsive part was

then taken to be the difference from the full potential

$$u_{\text{rep}}(r) = u(r) - u_{\text{att,fit}}(r). \quad (5.2)$$

In a second step, the repulsive potential up to the first RDF peak was fitted with the functional form  $u_{\text{rep,fit}}(r) = \frac{C_{\text{rep}}}{r^n}$ . For both fittings the residual was weighted with the value of the RDF, to ensure that we fit the relevant regions of the potential. If we assume the potential to be of LJ form, the repulsive part should be proportional to  $r^{-12}$ . In table 5.4 we present the parameters for fits of the optimized potentials as described in the main paper. We note that a one-step fit of the potentials with the form  $\frac{C_{\text{rep}}}{r^n} - \frac{C_6}{r^6}$  is less stable but leads to a better overall representation of the IMC potential and slightly smaller exponents.

Table 5.4: Fitting parameters resulting from first fitting the attractive tail of the pair potential and then the repulsive region of the remaining potential.  $u_{\text{rep,fit}}(r) = \frac{C_{\text{rep}}}{r^n}$  and  $u_{\text{att,fit}}(r) = -\frac{C_6}{r^6}$ .

		$C_{\text{rep}}$ in kJ/mol $\times$ nm <sup>n</sup>	$n$	$C_6$ in kJ/mol $\times$ nm <sup>6</sup>
HMN IMC	O - Li	$1.23 \times 10^{-3}$	6.9	$4.00 \times 10^{-3}$
	O - Na	$2.14 \times 10^{-4}$	8.4	$4.23 \times 10^{-3}$
	O - K	$8.51 \times 10^{-5}$	9.8	$7.08 \times 10^{-3}$
	O - Ca	$6.37 \times 10^{-3}$	6.9	$1.52 \times 10^{-2}$
	O - Cl	$7.22 \times 10^{-5}$	10.7	$3.76 \times 10^{-3}$
ECC IMC	O - Li	$1.64 \times 10^{-5}$	8.9	$1.18 \times 10^{-3}$
	O - Na	$5.33 \times 10^{-6}$	10.4	$1.62 \times 10^{-3}$
	O - K	$1.52 \times 10^{-6}$	12.3	$1.86 \times 10^{-3}$
	O - Ca	$5.28 \times 10^{-6}$	10.9	$3.90 \times 10^{-3}$
	O - Cl	$2.53 \times 10^{-6}$	13.2	$4.61 \times 10^{-3}$

Figure 5.6 shows the IMC potentials along with a fit of the attractive tail. There are some regions with unsmooth behavior. This originates from poor statistics at distances in between the first and second solvation shell. It does not lead to unphysical behavior in the simulations.

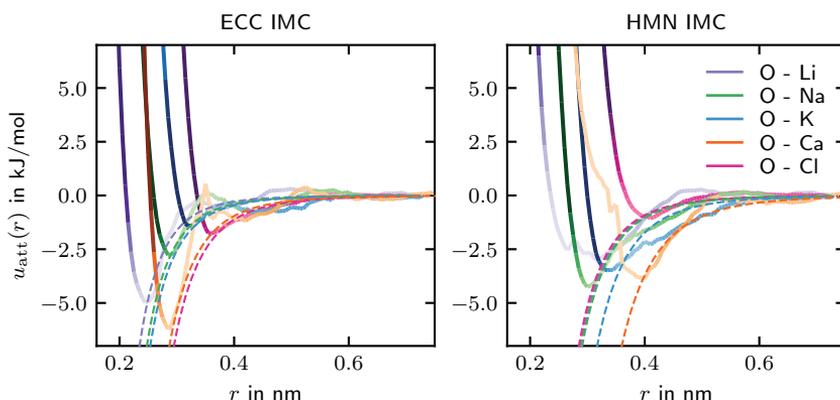


Figure 5.6: The obtained ion-oxygen potentials depicted as solid lines, shaded darker for higher values of the associated RDF. The fits of the attractive part are depicted as dashed lines.

### 5.2.5 LJ potentials and High Coulomb Attraction

Here we explain the argument for why the form of the repulsive part of the short-range potential is very important in electrolyte systems in detail. A Lennard-Jones potential has two parameters,  $\epsilon$  and  $\sigma$

$$u_{\text{LJ}}(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right]. \quad (5.3)$$

If two atoms carry charges  $q_1$  and  $q_2$  we add the Coulomb potential

$$u_{\text{E}}(r) = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r} \quad (5.4)$$

where  $\epsilon_0$  is the vacuum permittivity. If the charges are opposite, this pulls the atoms together into the repulsive part of the short-range potential. Take for example the attractive interaction between a sodium ion described by the OPLS force field and an SPC/E oxygen described using geometric combining rules. The minimum for their pair-potential  $u_{\text{LJ}} + u_{\text{E}}$  is found at 0.216 nm and therefore far in the repulsive regime of the LJ potential ( $\sigma_{\text{O-Na}} = 0.325$  nm).

Furthermore, the  $r^{-6}$  LJ term of the cation-oxygen interaction can become negligible which we demonstrate here briefly. For  $r < \sigma$  the attractive term is small relative to the repulsive term. For  $r > \sigma$  the attractive term tends to be small compared to the attraction from the Coulomb potential, depending on  $\epsilon$ ,  $\sigma$ , and the charges. As for small

$\sigma$  and large charges the Coulomb interaction is very strong we can expect this to be true, especially for small or multivalent ions. Given those two conditions, the  $r^{-6}$  term has little overall influence and the repulsive LJ term determines the shape of the effective potential minimum. Its form  $u_{\text{LJ,rep}} = 4\epsilon(\sigma/r)^{12}$  reveals that its two parameters ( $\epsilon$  and  $\sigma$ ) are then dependent. This effect is, to some extent, compensated by the water's hydrogen atoms because their positive charges add electrostatic repulsion between sodium and the water molecule. Still, due to the preferred orientation of water molecules in the solvation shell of cations, the effect is prevalent.

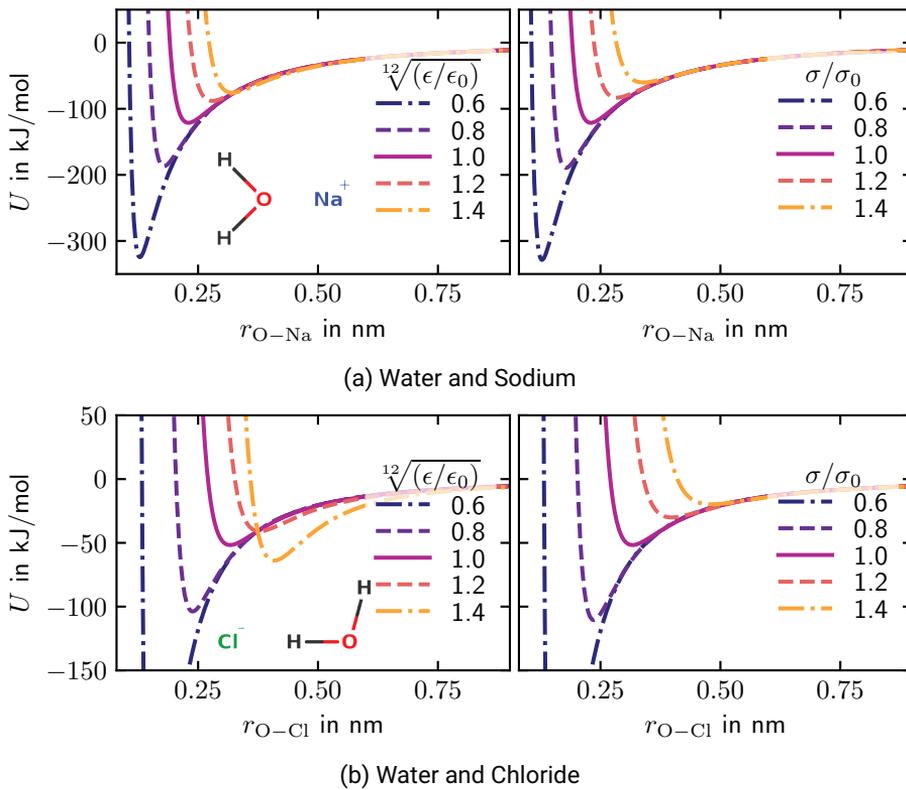


Figure 5.7: The total potential energy of a) a sodium or b) a chloride ion and a water molecule in a vacuum as a function of distance. The water molecule is oriented as shown in the inset. In the left and right graph,  $\epsilon$  and  $\sigma$  for the oxygen-ion interaction are varied, respectively, while the other is kept constant. The reference values a)  $\epsilon_0 = 0.0868 \text{ kJ mol}^{-1}$ ,  $\sigma_0 = 0.325 \text{ nm}$  and b)  $\epsilon_0 = 0.566 \text{ kJ mol}^{-1}$ ,  $\sigma_0 = 0.374 \text{ nm}$  refer to the OPLS+SPC/E parameters.

The effect is demonstrated in fig. 5.7, where we show how the total potential of a water molecule and a single ion in a vacuum varies with  $\sigma$  and  $\epsilon$ . As expected from the dominance of the repulsive term, a change in  $\sigma$  by a factor of e.g. 0.8 leads to a similar curve as scaling  $\epsilon$  with  $0.8^{12} \approx 0.06872$ . For anions, a similar argument can be made, but the situation is slightly different since the water molecules orient such that hydrogen atoms point towards the negative charge. For very high value of  $\epsilon$  ( $\epsilon = 1.4^{12}\epsilon_0$ ) we see the assumption of the LJ attractive term being smaller than the Coulomb attraction begins to break down as the two graphs differ slightly. This happens faster than for the sodium case for two reasons: (i) hydrogen has only half the charge of oxygen and (ii)  $\epsilon_0$  is already much higher for the Cl-O LJ interaction.

In the liquid phase, the more flexible orientation of the water molecules and the influence from other water molecules in the hydration shell can be expected to moderately tone down this effect. From the perspective of optimizing force fields, the dependence of  $\sigma$  and  $\epsilon$  is unfortunate, as one wants to have the parameters to be as independent as possible. However, it does not automatically mean that LJ is to “hard” or “soft”, just that it is not very flexible in modeling the real potential when high charges are involved.

## 5.2.6 Ion Vibrational Density of States

The formula for the DOS of a type of ion is

$$DOS(\nu) = \frac{2m}{k_B T N \tau} \sum_{j=1}^N \sum_{d \in \{x,y,z\}} |\mathcal{F}(v_{jd}(t))|^2. \quad (5.5)$$

Here,  $m$  is the ion mass,  $k_B$  is the Boltzmann constant,  $N$  is the number of ions,  $v_{jd}$  is the  $d$ -component (x, y, or z) of the velocity of ion  $j$ , and  $\mathcal{F}$  denotes the Fourier transformation. The time  $\tau$  is the length of the trajectory over which  $v_{jd}$  is defined. The sum runs only over the ions of one type and the velocities of other ions and water molecules do not go directly into this property. The DOS were calculated using our code available at <https://github.com/marvinbernhardt/dos-calc>. Figure 5.8 shows the density of states (DOS) of the ions. Due to the normalization in equation 5.5 the spectra have an integral of three, the number of translational degrees of freedom per ion. For detecting the peaks of the DOS we fitted them with the sum of two or three Lorentzian functions

$$DOS^{\text{fit}}(\nu) = \sum_{i=1}^3 \frac{a_i}{\pi} \frac{\gamma_i}{(\nu - \nu_i^0)^2 + \gamma_i^2}. \quad (5.6)$$

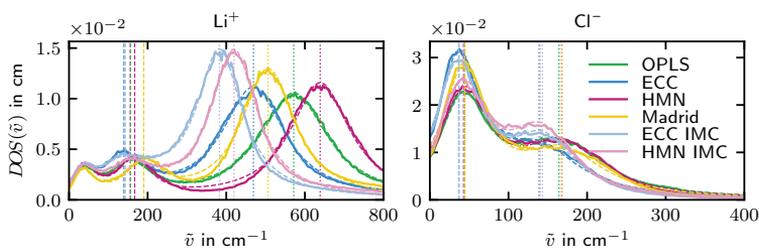
Here,  $a_i$  is the weighting of Lorentzian  $i$ ,  $\gamma_i$  determines its width, and  $\nu_i^0$  is its peak frequency. For  $K^+$ , the right peak is smeared out and we did not compare its frequency with experimental values. For  $Ca^+$ , the highest frequency peak is clearly visible, but we did not find experimental values to compare it to. Instead we use the second highest

---

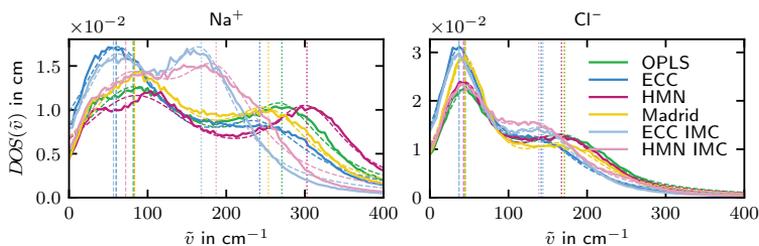
---

frequency peak of  $\text{Ca}^{2+}$  for which experimental data is available and compare that peak in the main text (fig. 5.2a).

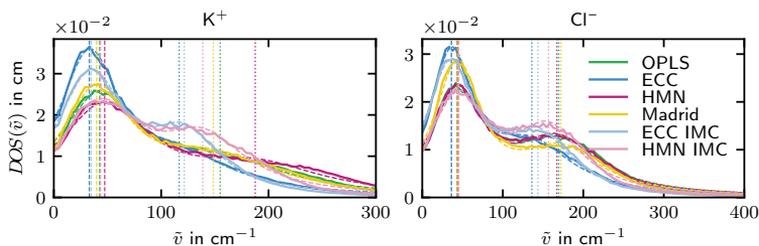
To ensure that our interpretation of the highest frequency peak in the DOS as the rattling motion of the ion is correct, we also applied the harmonic approximation to the cation-oxygen potential of mean force (PMF). In Figure 5.9 we show the PMF  $w(r) = -k_{\text{B}}T \log g(r)$  for NaCl and LiCl solutions. Also shown is a quadratic fit to the lowest three points of the PMF with the form  $w_{\text{fit}}(r) = ar^2 + br + c$ . From this a frequency can be obtained by  $f_{\text{pmf}} = \frac{1}{2\pi} \sqrt{\frac{2a}{\mu}}$  where  $\mu$  is the reduced mass of the cation and a water molecule. The frequencies are shown in fig. 5.2a in the main text. They match the highest frequency peak of the DOS and thereby indicate that the peak is caused by the ion rattling against the water of its hydration shell.



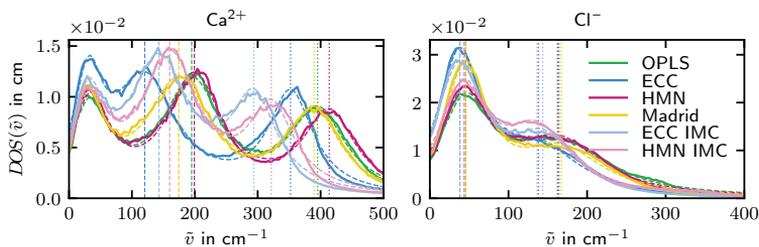
(a) LiCl



(b) NaCl



(c) KCl



(d) CaCl<sub>2</sub>

Figure 5.8: The Density of States of the ions in different salt solutions are shown as solid lines. The spectra are fitted with the sum of two ( $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Cl}^-$ ) or three (others) Lorentzian functions and are depicted as dashed lines. Vertical dotted and dashed lines indicate the highest and second highest frequency peaks taken from the fit parameters, respectively.

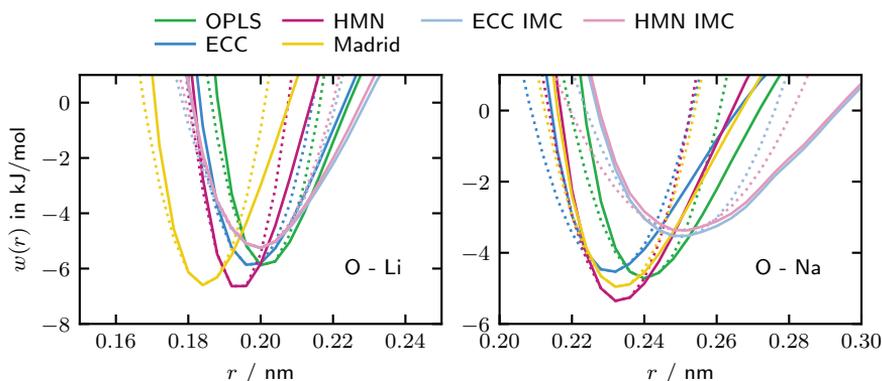


Figure 5.9: The cation-oxygen PMF minimum of NaCl and LiCl with six different force fields. The solids lines are the PMF as obtained by inverting the RDF which is shown in fig. 5.11. The dotted lines of the same color are the quadratic fit to the PMF minimum.

### 5.2.7 2PT Model Entropy

Solvation entropies of the salts in water by the 2PT model were not calculated directly but based on solvation entropies of the LJ models (ECC and HMN) obtained with the TI method (see next section). The 2PT model gives estimates for the total entropy of a system. In order to estimate the solvation entropy of the IMC models,  $\Delta S_{\text{solv}}(\text{IMC})$ , we use

$$\Delta S_{\text{solv}}(\text{IMC}) = \Delta S_{\text{solv,TI}}(\text{LJ}) + (S_{2\text{PT}}(\text{IMC}) - S_{2\text{PT}}(\text{LJ})) \quad (5.7)$$

where  $\Delta S_{\text{solv}}$  represents the sum of the individual cation and anion hydration entropies. 2PT entropies were calculated from the translational (trn), rotational (rot), and vibrational DOS of the water molecules and ions.[196] The idea of the 2PT model is to use a frequency dependent formula for the entropy and integrate it with the DOS to obtain the total entropy. The translational and rotational DOS are each split into two contributions, one diffusive (dif) and one oscillating (osc). The diffusive DOS is multiplied with the entropy of a hard-sphere  $S_{\text{HS}}$  gas (rigid rotator entropy  $S_{\text{RR}}$  for rotation). The oscillating DOS is multiplied with the entropy of a quantum harmonic oscillator  $S_{\text{HO}}(\nu)$  which is a function of the frequency. By integration over all frequencies, summation over all species and degrees of freedom, and adding a term for the mixing entropy an approximation of the total entropy of a system is obtained. The ions used in this work are all monoatomic and therefore have no rotational or vibrational degrees of freedom. Since the water models used are rigid, there are no vibrational degrees of freedom in our systems. We obtain the

following expression for the 2PT entropy of an aqueous electrolyte solution system

$$\begin{aligned}
S_{2PT} = & \sum_{h \in M} N_h \left[ \int_0^\infty S_{HS} DOS_h^{\text{trn,dif}}(\nu) d\nu + N_h \int_0^\infty S_{HO}(\nu) DOS_h^{\text{trn,osc}}(\nu) d\nu \right] \\
& + N_{H_2O} \left[ \int_0^\infty S_{RR} DOS_{H_2O}^{\text{rot,dif}}(\nu) d\nu + \int_0^\infty S_{HO}(\nu) DOS_{H_2O}^{\text{rot,osc}}(\nu) d\nu \right] \quad (5.8) \\
& - k_B \sum_{h \in M} N_h \ln x_h \\
\text{with } M = & \{H_2O, \text{Cation}, Cl^-\}.
\end{aligned}$$

Here,  $N_h$  and  $x_h$  are the count and mole fraction of molecule/ion type  $h$ , respectively. In order to be used in equation 5.7, the 2PT entropy of the system was normalized to a single unit of the salt. For the definition of the entropy functions and the method for obtaining the separated DOS we refer to the original paper.[196]

## 5.2.8 Thermodynamic Integration

The solvation free energies of the LJ models were calculated by coupling a single ion into a box of 506 water molecules using thermodynamic integration. First, the short-range interactions were turned on along a linear path of 9 intermediate Hamiltonians using soft core potentials for the LJ coupling.[205] Then the electrostatic interactions are turned on linearly over 9 intermediate steps. This is done separately for cation and anion by simulating a single ion in 506 water molecules. Each step is equilibrated for 400 ps and run for 400 ps in the NPT ensemble with the same settings as for the bulk simulations. The solvation free energy was then determined using standard trapezoidal integration rule of the average derivative of the potential with respect to the coupling parameter  $\lambda$ . Errors are estimated by the standard deviation of five independent runs.

To determine the solvation free energy of the IMC force fields the integration was performed with the respective LJ model as reference. The alchemical transformation switches the LJ potential off and the tabulated potential on, both linearly over 9 intermediate steps.

Several corrections have to be applied to the calculated solvation free energy from TI.[21] The finite size correction accounts for the self interaction of the ion and the orientational polarization of the solvent due to its periodic images.[206] It evaluates to

$$\Delta G_{fs} = \frac{N_A q^2}{4\pi\epsilon_0} \left[ \frac{-\xi_{ew}}{2\epsilon_r L} + \left( 1 - \frac{1}{\epsilon_r} \right) \left( \frac{2\pi R^2}{3L^3} - \frac{8\pi^2 R^5}{45L^6} \right) \right]. \quad (5.9)$$

Here,  $q$  is the charge of the ion,  $\xi_{ew} = -2.837297$  is a constant of the Wigner potential for a cubic box,  $L$  is the box length, and  $\epsilon_r = 71$  is the relative permittivity of the SPC/E

solvent, and  $R$  is the size of the ion, taken as the first peak of the oxygen-ion RDF. The pressure correction accounts for the ion in the uncoupled state referring to a compressed ideal gas, but the experiment takes the ideal gas at atmospheric pressure as reference. Its value is

$$\Delta G_{\text{press}} = N_A k_B T \log \frac{p_1}{p_0} = 8.0 \text{ kJ mol}^{-1} \quad (5.10)$$

Here,  $p_0 = 1 \text{ atm}$  and  $p_1 = 24.94 \text{ bar}$  is the pressure that compresses an ideal gas to  $1 \text{ mol L}^{-1}$  at 300 K. The transfer of the ion across the gas-liquid interface is also connected to an energy term that comes from the electrostatic potential drop across the air-water interface. Since we only look at the combined values of salt pairs (not at single ions) it cancels out. For force fields with scaled charges, the transfer from gas to liquid a configuration independent polarization term has to be added.[17] It represents the attractive energy between the charges in the system and the induced electronic dipoles in the solvent. Since this effect is only accounted for in a mean-field way in the ECC model and not explicitly in the Hamiltonian, it has to be accounted for when transferring an ion from gas to liquid phase. The correction is given by a Born-type expression:

$$\Delta G_{\text{pol}} = -\frac{1}{4\pi\epsilon_0} \left(1 - \frac{1}{\epsilon_{\text{el}}}\right) \frac{q_u^2}{2R_{\text{ion}}}. \quad (5.11)$$

Here  $\epsilon_{\text{el}}$  is the assumed electronic part of the permittivity of water, which is the inverse square root of the charge scaling factor and  $q_u$  is the unscaled charge of the ion. The ionic size  $R_{\text{ion}}$  is a parameter that should represent the boundary at which the assumed electronic continuum of the solvent molecules starts. The role of the radius or the boundary in the ECC model has been discussed critically for cavities in proteins.[207]. For the alchemical change from LJ to IMC potential we apply no corrections.

In fig. 5.10 we show the solvation free energies of the salts with all models. Note that while we report the combined values for the cations and anions, the reference state is still the isolated ions in the gas state, not the crystalline salt. For the models with scaled charges, we show the value of the free energy without polarization correction and three values based on different radii including the polarization correction. It becomes evident, that the solvation free energy of the OPLS and HMN model well reproduces the experimental value for all salts. Our HMN IMC model shows slightly higher values compared with the HMN model but is still close to the experimental value. The ECC, ECC IMC, and Madrid model show values that depend strongly on the choice of the ionic radius. This is unfortunate for these models, as comparisons to experimental values depend on a parameter. The two more sensible choices for the ionic radius are (i) taking a literature value and (ii) deriving it from the ion-oxygen and oxygen-oxygen RDF. For both of these choices, the values for the solvation free energy are similar but much too low for all salts. In the ECC papers on ions in water, the solvation free energy has never been explicitly calculated.[18, 20, 167, 190, 191, 209]

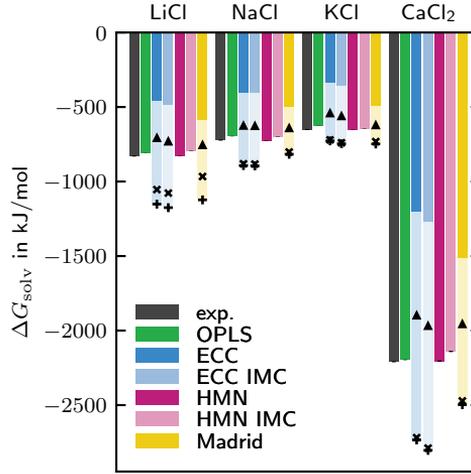


Figure 5.10: Solvation free energies of the salts. For all models with scaled charges, the opaque bars show the value without the polarization correction. The values obtained by including the polarization correction (Equation 5.11) for those models are shown as symbols:  $\times$  with  $R$  as the literature value ionic radius from Marcus[208], and  $+$  with  $R$  as the ionic radius from the first peak in the ion-oxygen RDF minus the water radius of 0.138 nm.

The solvation entropy of the LJ models is determined from

$$\Delta S_{\text{solv}} = \frac{\Delta H_{\text{solv}} - \Delta G_{\text{solv}}}{T}. \quad (5.12)$$

The enthalpy difference  $\Delta H_{\text{solv}}$  is obtained by simulating the two endpoints for 400 ns. At the uncoupled endpoint, instead of simulating the ion in an ideal gas state, it was not simulated. Its contribution of  $\frac{5}{2}k_B T$  that arises from its translational energy was subtracted from the enthalpy difference manually to get  $\Delta H_{\text{solv}}$  in equation 5.12. For the solvation enthalpy, an additional enthalpy correction of  $k_B T^2 \alpha$ , with the thermal expansion coefficient  $\alpha$ , has to be subtracted from the enthalpy difference.[210, 211] In practice we found it to be smaller than  $0.5 \text{ kJ mol}^{-1}$  and therefore not relevant.

The corrections to the free energies described above lead to analogous corrections to the solvation entropy by their temperature dependence. The finite size correction becomes

$$\Delta S_{\text{fs}} = -\frac{N_A q^2}{4\pi\epsilon_0} \frac{1}{\epsilon_r^2} \frac{\partial \epsilon_r}{\partial T} \left[ \frac{-\xi_{\text{ew}}}{2L} - \frac{2\pi R^2}{3L^3} + \frac{8\pi^2 R^5}{45L^6} \right]. \quad (5.13)$$

The temperature derivative of the relative permittivity is  $-0.3631 \text{ K}^{-1}$ . The pressure

correction becomes

$$\Delta S_{\text{press}} = -N_A k_B \log \frac{p_1}{p_0} = -26.6 \text{ J mol}^{-1} \text{ K}^{-1} \quad (5.14)$$

For the IMC models the entropy of the LJ models was taken as reference

$$\Delta S_{\text{solv}}(\text{IMC}) = \Delta S_{\text{solv}}(\text{LJ}) + \frac{\Delta H(\text{LJ} \rightarrow \text{IMC}) - \Delta G(\text{LJ} \rightarrow \text{IMC})}{T}. \quad (5.15)$$

The solvation entropies of the salts are obtained from the sum of the individual cation and anion contribution. We calculated them for ECC, ECC IMC, HMN, and HMN IMC and report them in fig. 5.2b. We reproduce the solvation free energies and solvation entropies for the HMN model.[21, 172] We however see a much slower convergence of  $\Delta H$  which we attribute to using a different thermostat (Horinek et al.: Berendsen, this work: Nosé-Hoover).

## 5.2.9 Other Properties of Simulated Electrolytes

In the following, we show all the measured properties of the six different force fields. The RMSD  $\Delta_x$  for judging the quality is defined as follows

$$\Delta_x^2 = \sum_{s \in \{\text{LiCl, NaCl, KCl, CaCl}_2\}} \sum_{r \in \{0.005, 0.01, 0.02, 0.03, 0.04, 0.05\}} (x_{sr} - x_{sr, \text{exp}})^2 \quad (5.16)$$

Here,  $s$  is an index for the salt,  $r$  is the molar mixing ratio, and  $x$  is one of  $\rho$  (density),  $D(\text{H}_2\text{O})/D_0(\text{H}_2\text{O})$  (water self diffusion), and  $\phi$  (osmotic coefficient). We also simulated salt solutions with  $r = 0.1$  but often there is no experimental value for this high concentration, so we excluded it here. For the experimental data, if there was no data for the precise concentration or temperature, we used linear interpolation between existing data points. For the quality of the ion-oxygen RDF we compare with the AIMD results.  $\Delta_{\text{RDF}}$  is defined as

$$\Delta_{\text{RDF}}^2 = \sum_{s \in \{\text{LiCl, NaCl, KCl, CaCl}_2\}} \sum_{i \in \{\text{O-cat., O-an.}\}} \frac{w_i}{T_{\text{max}}} \int_0^{T_{\text{max}}} (g_{si}(r) - g_{si, \text{AIMD}}(r))^2 dr \quad (5.17)$$

Indices  $s$  and  $i$  count over the four different salts and the two separate interactions, respectively. Note that here  $r$  is the distance, not the mixing ratio. To not overweight the O-Cl RDF, which is present in all four salts we use a weighting factor  $w_i$  which is one for cation-oxygen and 0.25 for chloride-oxygen. The RDF was calculated at a concentration of 50 formula units of salt per 5000 water molecules.

## RDF

The RDFs were calculated at a molar mixing ratio of  $r = 0.01$ . See all RDF data in fig. 5.11.

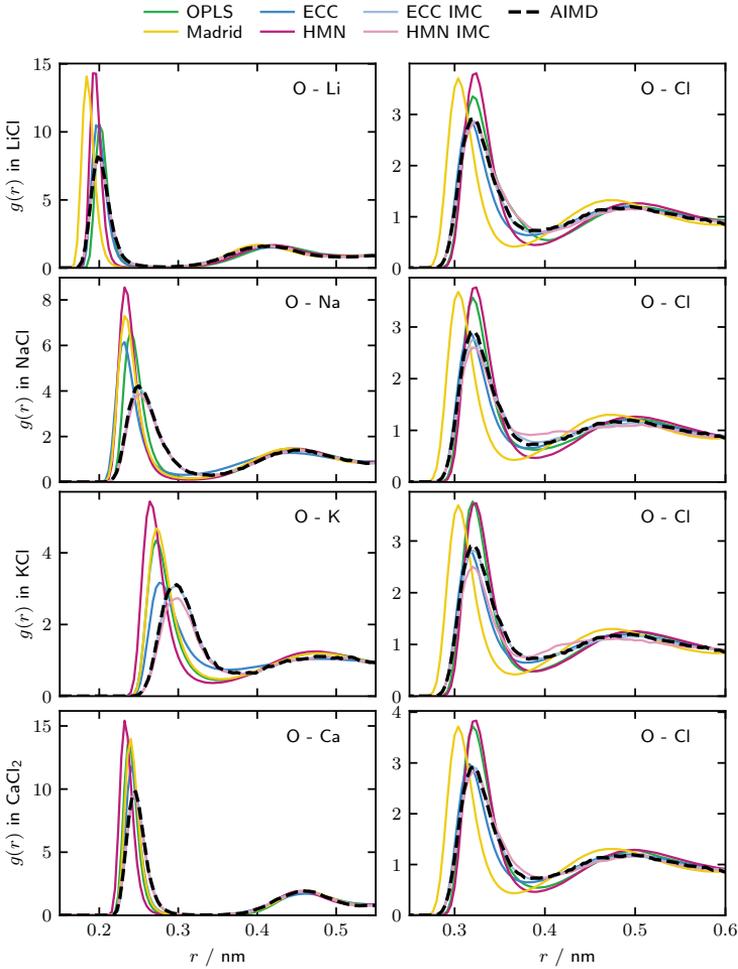


Figure 5.11: The RDF of the oxygen-cation and oxygen-chloride distance for eight different force fields and four different salts.

## Density

The solution density data are shown in fig. 5.12 as a function of the salt concentration. Those were calculated from 0.4 ns NPT runs with a Parinello-Rahman barostat at 1 bar.

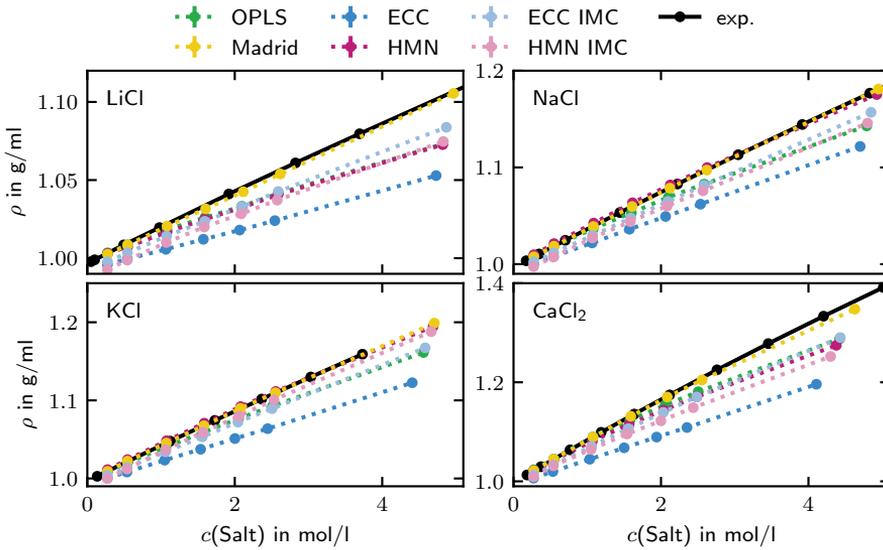


Figure 5.12: The densities obtained from NPT runs of the different force fields. Experimental data is taken from Perry's and was linearly interpolated to a temperature of 300 K.[212]

## Water Diffusion

Figure 5.13 shows the water diffusion coefficient as a function of salt concentration normalized by the water diffusion self-diffusion coefficient ( $D_0$ ) at zero salt concentration. A water mean-square displacement of 400 ps length was calculated for each system from the NPT runs. The diffusion coefficient ( $D$ ) was obtained from a linear fit from 40 ps to 360 ps. It was then corrected for errors from finite size effects[213] by

$$D_{\text{corr}} = D_{\text{MD}} + \frac{\xi k_B T}{6\pi\eta L}. \quad (5.18)$$

The viscosity  $\eta$  is 0.729 mPa s for SPC/E and 0.855 mPa s for TIP4P/2005[214] and is taken to be approximately constant with the ion concentration. The dimensionless constant  $\xi$

for cubic boxes has a value of 2.83729. We note that the correction has little effect on the quotient of two diffusivities obtained from roughly similar system sizes.

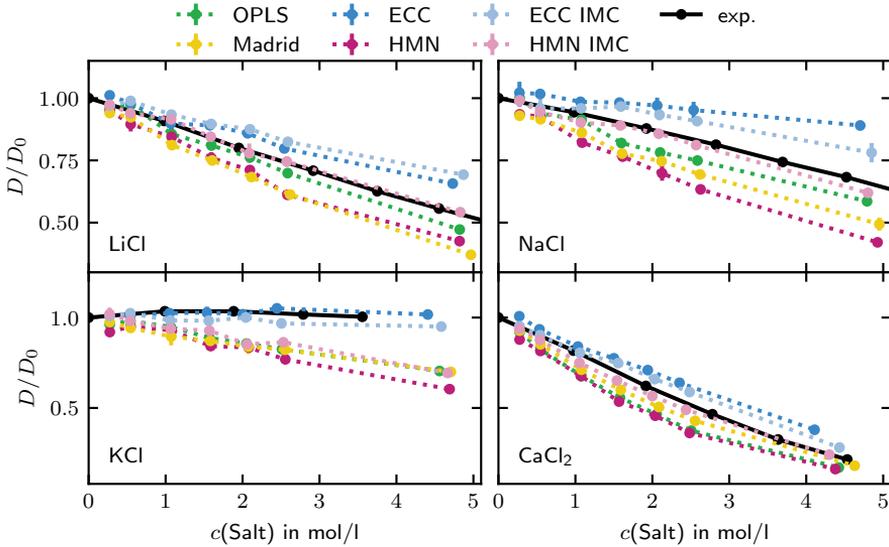


Figure 5.13: The diffusion coefficient of water divided by the pure water self-diffusion coefficient obtained from NPT runs of the different force fields is shown. Experimental data is from NMR measurements by Müller and Hertz at 298.15 K.[215]

### Osmotic Coefficient

Figure 5.14 shows the calculated osmotic coefficients. We obtained the osmotic coefficient using the OPAS method, which involves having an elongated box with two regions.[140] These regions are separated by semi-permeable half-harmonic potential walls ( $k = 4000 \text{ kJ nm}^{-2}$ ) only permeable for the water. In a pre-run, the osmotic pressure is calculated for some fixed total pressure. From that estimate of  $\Pi$ , we fix the total pressure such that the pressure of the water slab has a pressure of approx. 1 bar. The final production run is 10 ns long. Since GROMACS does not export the force on/by the semi-permeable wall  $f_{\text{wall}}$ , we use our code for its calculation from the trajectory, available at <https://github.com/marvinbernhardt/wallforce>. The osmotic pressure is calculated

as

$$\Pi = \frac{\langle f_{\text{wall}} \rangle}{A_{xy}} \quad (5.19)$$

where  $A_{xy}$  is the area of the wall. The osmotic coefficient  $\phi$  is given by

$$\phi = \frac{\Pi}{icRT} \quad (5.20)$$

with the idealized (full dissociation) van 't Hoff factor  $i$  and concentration  $c$ .

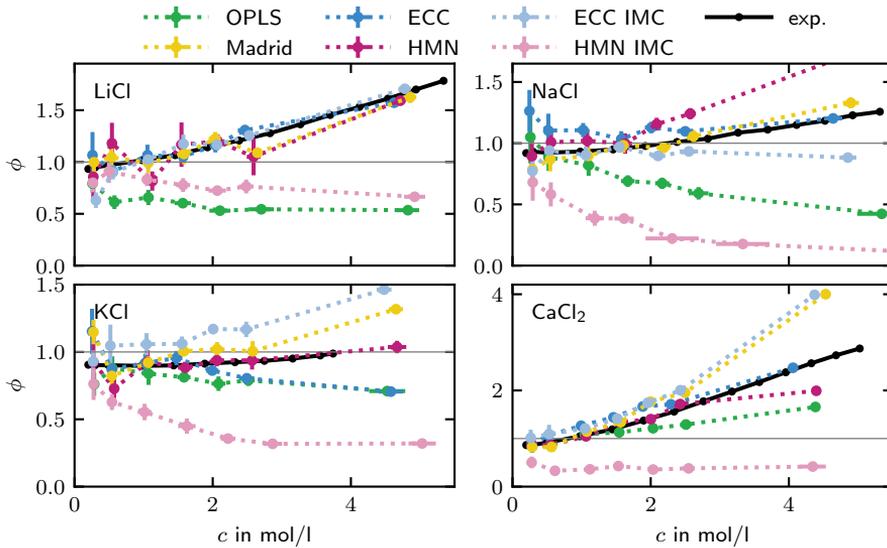


Figure 5.14: Osmotic coefficients of four salts and six different force fields. Experimental data determined with a hygrometric method by Guendouzi et al. has been adapted using experimental densities from the CRC Handbook.[216, 217]

---

---

## 6 Application of the 2PT model to understanding entropy change in molecular coarse-graining

---

### Abstract

The effects of coarse-graining on the thermodynamic properties of molecular liquids are studied by computing the standard molar entropies of chloroform, 1,4-dioxane and neopentane using the two-phase thermodynamic (2PT) model. From the trajectory of molecular dynamics simulations, the entropy associated with changing the level of resolution is measured. The entropic contribution coming from the choice of the mapping and from the coarse-grained force field are decoupled. The entropy of the eliminated degrees of freedom is estimated from the rerun of the mapped trajectory, while the additional entropy variation, tied to the coarse-grained force field itself, is quantified by subtraction. In this work the 2PT model is used as a diagnostic tool to analyze how different mappings schemes affect the dynamic modes of the molecules and to quantify the change in entropy. Good agreement between the target density of states of the mapped trajectory and the coarse-grained distribution of normal modes is found. The entropy change upon coarse-graining is discussed in terms of vibrational, rotational and translational contributions.

### 6.1 Introduction

In the field of systematic coarse-graining several approaches for deriving coarse-grained (CG) potentials are established or part of ongoing research.[27, 40, 82, 218] For all bottom up methods the first step is the choice of a mapping scheme that determines which atoms each CG bead is representing. However, the role of the mapping scheme in the effects on dynamic modes and the absolute entropy of a CG system are not sufficiently studied. Here we use a method which allows us to analyze effects of mapping on dynamic modes of molecular fluids in relation to changes in the absolute entropy. That connection is given by the two-phase thermodynamic (2PT) model which calculates the translational,

---

---

rotational, and vibrational contributions to the absolute entropy from the corresponding contributions of the (vibrational) density of states (DoS).[195]

The acceleration of dynamics in CG models is a known effect of the softening of the interactions between particles and the change in the surface area available for collisions.[219] Efforts to correct the dynamics of CG force fields are made introducing Markovian dissipative forces.[220–224]. The success of those approaches depends mainly on the assumed time scale separation of the CG and removed degrees of freedom. While we are not treating dissipative forces in this study, time scale separation will be part of our analysis with an eye to the choice of mapping schemes. A connection of dynamics to structure and thermodynamics is provided by the excess entropy scaling relations described by Rosenfeld and others who found that dynamical properties such as diffusivity correlate well with the excess entropy of a fluid.[225, 226] This connection can be used to predict the speed up of dynamics in polymer systems upon coarse-graining.[90] However, the changes in dynamics of CG systems are more complex than a simple speed up, especially in complex systems, where both friction and barrier crossings are important.[227] In general, reducing the number of particles that represent a molecule does not only reduce the effective friction with other molecules and barrier heights in the free energy landscape, it can also affect its shape, symmetry, moment of inertia, and internal degrees of freedom (DoF). Special mappings with virtual sites have been used to retain the molecular symmetry.[228] Mapping a molecule to multiple beads introduces new bonded interactions, which are often parametrized to match the internal structure, not the dynamics, of the reference molecule.[29, 229] For proteins coarse-graining affects internal dynamics of the molecules as low frequency modes increase in weight and fast processes are eliminated.[31] The connection of the free energy landscape with the transition dynamics in proteins has been studied in great detail revealing relationships between structure and (long time-scale) dynamics if the steric effects are represented sufficiently in the coarse-grained model.[230] For liquid molecular systems a helpful tool for the examination of dynamics is the DoS. The DoS function or power spectrum of a system is proportional to the averaged squared absolute of the Fourier-transform of the velocities. Its value at zero frequency is related to the diffusion coefficient, and the DoS is proportional to the Fourier transform of the velocity autocorrelation function. A separation of the (atomic) DoS into translational, rotational, and vibrational (molecular) contributions is possible.[196]. Naturally, vibrational modes separate into peaks in the spectrum, providing us a more detailed view on internal dynamics. These are important features which allows us to investigate the dynamics of different motions in a unified way. Thereby, it is easier to understand which dynamics happen on a similar timescale and which are suppressed in a CG model.

Along the effects of coarse-graining on the dynamic properties of the system, implications on thermodynamic properties are discussed in this paper. Generally, properties like energy, entropy, hydration free energy, heat capacity, pressure, and thermal expansion coefficient are not conserved upon coarse-graining but can be, depending on the

---

---

method used, in qualitative agreement[231–233], or can be accounted for by introducing correction terms[64]. Kmiecik et al. point out that some CG models reproduce the free energy differences of their reference systems but have a different enthalpy entropy balance. [234] However, the representability of physical properties does not only depend on the choice of the CG force field. Already the application of the mapping operator reduces the dimensionality and thereby the number of microstates of the system. Entropy is a measure of the number of accessible microstates and therefore a key property to understand the effects of different mapping schemes. In previous works, the loss in configurational entropy upon coarse-graining has been measured for phospholipids and hydrocarbons using Schlitter’s quasi-harmonic approximation.[231, 235] Baron et al. showed that the loss in entropy is higher for tail-fragments of long linear alkane chains, as the terminal-part of the molecule is more flexible.[231] For protein domains the impact of resolution upon entropy has been quantitatively assessed by Foley et al. using a Gaussian Network Model.[31] They show that entropy loss is not scaling linearly with coarsening and that the information content per site is at its highest for intermediate value of coarse-graining.[31] In both articles [31, 231] entropy is also discussed in terms of mapping entropy, which depends solely on the mapping operator. By comparing the probability distribution of mapped and unmapped atomistic configurations, the entropy difference between the all-atom (AA) and CG model is computed, providing information on the mapping scheme. The mapping entropy measures the effect of the mapping on the information loss, irrespective on the way the CG interactions are described or derived, and can guide the choice of the CG architecture. The concept of mapping entropy has been defined within the framework of statistical mechanics in a rigorous way.[56, 236] Additionally, an analytical estimate for the mapping entropy associated with changing the level of resolution of a polymer chain has been given by Guenza, using their analytical Integral Equation Coarse-Graining method.[126] The knowledge of the entropy of the system can be used as a strategy to optimize CG force fields. Also the 2PT method has been used before in the context of coarse-graining by Jin et al. to investigate the entropy loss when using single bead mapping for methanol and chloroform and to improve their temperature and composition transferability.[237] Shell and coworkers used relative entropy as a tool to derive bottom-up potentials by minimizing the difference between AA and CG probability distributions.[56, 84, 232]

For this work, three different CG representations of chloroform, 1,4-dioxane and neopentane are defined and the pertaining CG potentials derived. To obtain the non-bonded interaction of the CG model the Conditional Reversible Work (CRW) method, a bottom-up approach, is followed.[47, 238] The CG models are compared to the underlying AA system in terms of structural, dynamic and thermodynamic properties. The ability of the CG models to capture the AA reference is critically discussed. While structural properties are considered in section 6.4.1, the main focus of the work lies in the analysis of dynamic modes and standard molar entropies of our systems, sections 6.4.2 and 6.4.3. Spectra of the CG model and the mapped AA model are compared to the AA DoS, where

we put some emphasis on the aspect of time scale separation. We observe effects of coarse-graining on the internal and rotational dynamics of molecular liquids, that have drawn little attention in literature. The 2PT method yields us contributions of translation, rotation and vibration to the entropy, which makes it a well suited tool for a in-depth analysis. While the 2PT model is approximative, it has been used successfully used to estimate entropies of water in confinement[239] and molecular liquids[240]. The technique and the details of the implementation are discussed in section 6.2.1. In this work, the ability of the 2PT approach to capture the entropic change in moving from an AA to a CG description of the system is put to test. To discriminate between the entropy change due to a loss in resolution and the entropy variation that comes from the limitations of using an approximated potential, the mapping entropy of the system is considered.

## 6.2 Theoretical Background

### 6.2.1 2PT Model

The 2PT model is used to estimate the absolute entropy from MD and mapped trajectories. It works under the assumption of having the molar entropy of a pure molecular fluid  $S$  given as a sum of translational, rotational and vibrational contributions

$$S = S_{\text{trn}} + S_{\text{rot}} + S_{\text{vib}}. \quad (6.1)$$

Each of them is calculated from the DoS function of the respective DoF. The separability of the total DoS is a consequence of the general decomposition of the velocities of the atoms.

In the model the translational DoS function is the sum of two additive parts: the DoS of a hard sphere gas,  $DoS_{\text{trn}}^{\text{hs}}$ , with  $f_{\text{trn}}N$  particles and the DoS of  $(1 - f_{\text{trn}})3N$  independent quantum harmonic oscillators,  $DoS_{\text{trn}}^{\text{ho}}$ . The factor  $f_{\text{trn}}$  is a measure of the systems fluidicity, which approaches 1 for dilute gases and 0 for solids. The determination of  $f_{\text{trn}}$  from the diffusion coefficient is a key step in the 2PT approach and a general formula is derived from hard sphere theory.[195] The shape of the hard sphere gas DoS is entirely specified, once  $f_{\text{trn}}$  is known. By taking the difference to the total translational DoS, the contribution of the oscillators is obtained.

To obtain the entropy,  $DoS_{\text{trn}}^{\text{ho}}$  is integrated over all frequencies  $f$  with a weighting function  $W^{\text{ho}}$  that represents the entropy of a quantum harmonic oscillator. For the hard sphere fluid the weighting function  $W^{\text{hs}}$  is based on the Carnahan-Starling equation of state.[241] The latter weighting function is independent of the frequency. Summing the entropies of those two pseudo-subsystems gives the translational entropy

$$S_{\text{trn}} = Nk \int_0^{\infty} \left[ W^{\text{hs}} DoS_{\text{trn}}^{\text{hs}}(f) + W^{\text{ho}}(f) DoS_{\text{trn}}^{\text{ho}}(f) \right] df. \quad (6.2)$$

The formulas for the weighting functions and  $f_{\text{tm}}$  are reported in the original 2PT papers.[195, 196]

Since the rotational DoS function has a zero frequency contribution, which comes from the rotational diffusion of the molecules, it is also split in two additive contributions. In the 2PT model it is treated similar to the translational part. The same formula is used for the calculation of the rotational fluidicity factor  $f_{\text{rot}}$ . The diffusive part is modeled by a subsystem of  $f_{\text{rot}}N$  free rotors ( $DoS_{\text{rot}}^{\text{fr}}$ ), while the oscillating part, representing the librations of the molecules, is portrayed as independent quantum harmonic oscillators( $DoS_{\text{rot}}^{\text{ho}}$ ). Summing the contributions yields the rotational entropy

$$S_{\text{rot}} = Nk \int_0^{\infty} \left[ W^{\text{hs}} DoS_{\text{rot}}^{\text{fr}}(f) + W^{\text{ho}}(f) DoS_{\text{rot}}^{\text{ho}}(f) \right] df. \quad (6.3)$$

A separation of the vibrational DoS function becomes necessary if torsional rotation occurs in the system. It has been shown that the torsional entropy of flexible molecules can be calculated in the 2PT framework.[242] The systems studied in this paper do not show any zero frequency contributions in the vibrational DoS. Therefore, all internal motions of the molecules are treated as independent quantum harmonic oscillators with their entropy given by

$$S_{\text{vib}} = Nk \int_0^{\infty} W^{\text{ho}}(f) DoS_{\text{vib}}(f) df. \quad (6.4)$$

Entropies reported in this paper are calculated with the 2PT model as described in [196], but with three deviations.

First, for linear molecules the angular velocity can not simply be derived from  $\vec{\omega} = \mathbf{I}^{-1} \vec{L}$  since the moment of inertia tensor  $\mathbf{I}$  is singular ( $\vec{L}$  is the angular momentum). Instead, the angular velocity is obtained from solving the underdetermined system  $\mathbf{I}\vec{\omega} = \vec{L}$  with minimum angular momentum ( $\vec{L}$ ) norm. Also, for linear molecules it is not possible to consistently define the two principal axes perpendicular to the molecules axis. Therefore the basis of the angular velocity vector is left untransformed in the lab coordinate system. The 2PT model has been used for  $\text{CO}_2$  which is in principle linear, but can be tackled with the normal 2PT toolset, because it exhibits an angular shape most of the time during MD simulations.[243]

Secondly, as the equipartition of kinetic energy is not always perfectly met in a finite time MD simulation, we use a first order correction to the DoS function as described in [244]. The DoS is scaled, such that its integral equals the number of DoF per molecule (not per system).

Thirdly, the rotational molar DoS is calculated by

$$DoS_{\text{rot}}(f) = \frac{2}{k_{\text{B}}TN} \sum_i^N \sum_{d \in \{a,b,c\}} \left| \mathcal{F} \left( \sqrt{I_{i,d}} \omega_{i,d} \right) (f) \right|^2 \quad (6.5)$$

---

---

where  $\mathcal{F}$  denotes the Fourier transform over a finite period  $T$  normalized with  $\frac{1}{\sqrt{T}}$ ,  $\omega_{i,d}$  is the angular velocity around axis  $d$ , which is one of the three principal axes of rotation of molecule  $i$ , and  $I_{i,d}$  is the respective moment of inertia. The formula deviates from the formulation of the 2PT method, since the square root of the moment of inertia is inside the Fourier transform. We think this is more consistent, since the moment of inertia is time dependent for flexible molecules. It can be shown, that otherwise, the integral of the  $DoS_{\text{rot}}(f)$  does not necessarily equal the number of rotational DoF.[245]

For a more detailed view on the method the reader is referred to the original papers[195, 196].

We note that, besides the principal axis frame used for the angular velocity in the 2PT method, there exists the Eckart frame, which is defined such that the rotational-vibrational (Coriolis) coupling is minimal.[246, 247] For stiff molecules the two reference frames are identical. Studies show that the use of the Eckart frame gives (compared to lab frame) more consistent results for star polymers under shear flow [248] and small linear molecules in gas phase.[249] For azulene in carbon dioxide and xenon it is found that using the Eckart and principal axis frame results in the same rotational energies.[250] The molecules considered in this work are rather stiff and the rotational and vibrational modes are always well separated, see the DoS in section 6.4.2. We therefore use the principal axis frame but note that for more flexible molecules this might lead to artifacts caused by Coriolis coupling.

## 6.3 Simulation Details

Starting from AA calculations as reference, the CG potentials are derived in a bottom-up approach. In this section, the atomistic model is introduced and the procedure for obtaining the CG potentials summarized. If not otherwise specified, the settings of the CG simulations are the same employed for the AA ones.

### 6.3.1 Atomistic Model

Atomistic simulations of chloroform, 1,4-dioxane and neopentane are performed using the OPLS-AA force field, which is designed for the modeling of organic molecules in the liquid phase.[7, 251] A spherical cutoff of 12 Å is applied to both Lennard-Jones and electrostatic interactions. To compensate for the truncation of the Lennard-Jones potential a long-range correction is added to the energy and pressure.[76] The electrostatic interactions are computed using the particle-mesh Ewald algorithm[252, 253] with cubic interpolation and 36 or 40 grid points per dimension. In the OPLS-AA force field a scaling factor of 0.5 is applied to intramolecular 1-4 interactions and a geometric mixing rule is employed to describe Lennard-Jones cross-interactions. Trajectories are generated using GROMACS 2018.[139] Newton's equation of motions are solved using a leap-frog

---

---

integration scheme with a time step of 1 fs. The algorithm is complemented by periodic boundary conditions in x, y, and z direction. In order to maintain a constant average temperature, a Nosé–Hoover thermostat[254] is used, with a oscillation time of 5 ps. All systems are studied in the liquid phase at 1 bar, setting the target temperature to 300 K for chloroform and 1,4-dioxane and to 275 K for for neopentane. A Parrinello–Rahman barostat[255] with a oscillation time of 30 ps is chosen to control the pressure. Each system consists of 512 molecules. After 200 ps equilibration (with Berendsen thermostat and Berendsen barostat) a 1 ns production run in the NPT ensemble is performed. The velocities and positions of the atoms are collected every 4 fs of the production run and taken as input for the calculation of the DoS function.

### 6.3.2 Coarse-Grained Model

Similarly to the atomistic force fields, the CG force fields have to account for bonded and nonbonded interactions. The CRW method is used here to generate nonbonded potentials,[47] while for deriving the bonded interaction the Boltzmann inversion approach is followed.[29] CRW computes the interaction free energy between CG sites at the pair level. A set of constraint atomistic dynamics in vacuum is performed, in which the distance that separates two CG groups is fixed. From the forces on the constraints the pair potentials are derived. The details of the method have been described in earlier papers.[47, 256] Using CRW models it is possible to achieve transferability to different state points.[257] The potential obtained with the CRW method have also proven to be chemically transferable and have been successfully applied to the study of apolar[47], polar molecules[238], and surfaces[258]. In the present case, the potentials are derived in vacuum, through a series of computationally inexpensive 20 ns stochastic dynamics simulations with a friction constant of  $0.01 \text{ ps}^{-1}$ . Those are run in the LAMMPS molecular dynamics package[259]. The shortest distance considered during the sampling is  $3.2 \text{ \AA}$ , whereas the longest distance, which coincides with the cutoff of the nonbonded CG potential, is  $13 \text{ \AA}$ . The in-between points are equally spaced with a  $0.02 \text{ \AA}$  increment. The derived potentials are used in their tabulated form. Following the prescription of Deichmann et al.[238] for polar molecules, the electrostatic component is considered for the 1,4-dioxane united-atom (UA) representation. In all other cases it is treated implicitly.[47] The simulations of the CG molecules are performed in LAMMPS with a Verlet integrator and a time step of 1 fs. A Nose-Hoover thermostat with a relaxation time of 0.1 ps and a Martyna-Tobias-Klein barostat with a relaxation time of 1 ps.[260, 261]. Temperature and pressure are kept to the same values as for the atomistic simulations. Where electrostatic interactions are needed they are computed using the particle-particle particle-mesh solver[262] with polynomial interpolation of order 5 and 12 grid points per dimension.

## 6.4 Results and Discussion

### 6.4.1 Structure

We introduce different CG representations of chloroform, 1,4-dioxane and neopentane at multiple resolutions. Three levels of coarse-graining are considered: a united-atom model, in which hydrogen atoms are merged to the closest heavy atom, a dumbbell representation, with two particles separated by a bond, and a single bead model. In all representations the site of the CG beads coincides with the center of mass of the group of atoms. The nonbonded potentials for different molecules and mapping schemes are collectively shown in fig. 6.1.

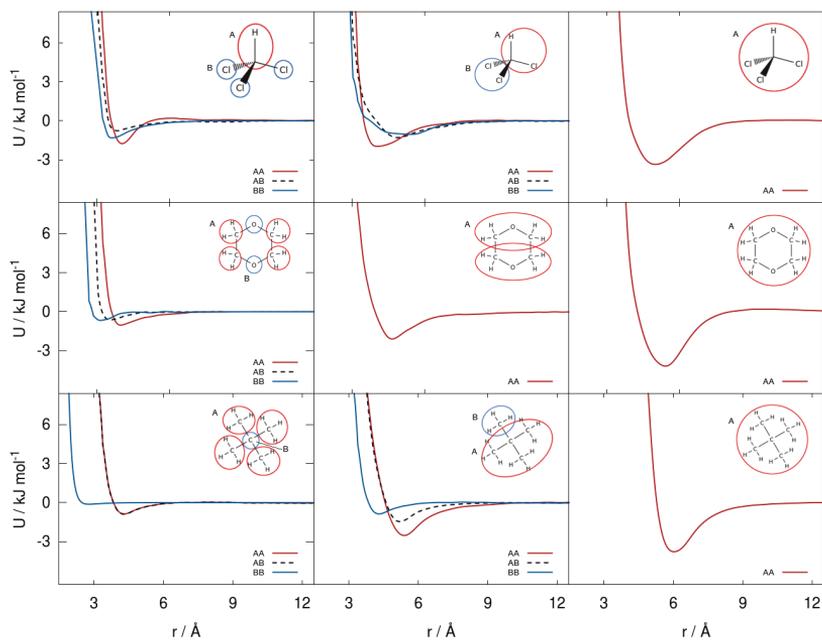


Figure 6.1: Overview of CRW-derived potentials for chloroform, neopentane and 1,4-dioxane. From top to bottom different mapping schemes are shown: united-atom, dumbbell, and single bead representation. A schematic view of the mapping schemes is given as an inset.

The CRW potentials are well-behaved with limited to none oscillation in the tail part of the curve. The plots are labeled with a schematic representation of the molecule, in which atoms belonging to a given CG site are circled. The different mapping schemes differ not only by the size of the beads, but also explore different symmetries. For example, in case of the dumbbell representation of neopentane, the molecule is not split into beads of equal size, but it is described as two asymmetric entities:  $\text{CH}_3$  and  $\text{C}_4\text{H}_9$ . It is not uncommon to have beads of different size used to portray a molecule.[263] In a polymer the side-chain beads and the backbone beads might have different sizes, as the choice of the coarse-graining scheme is often driven by chemical intuition.

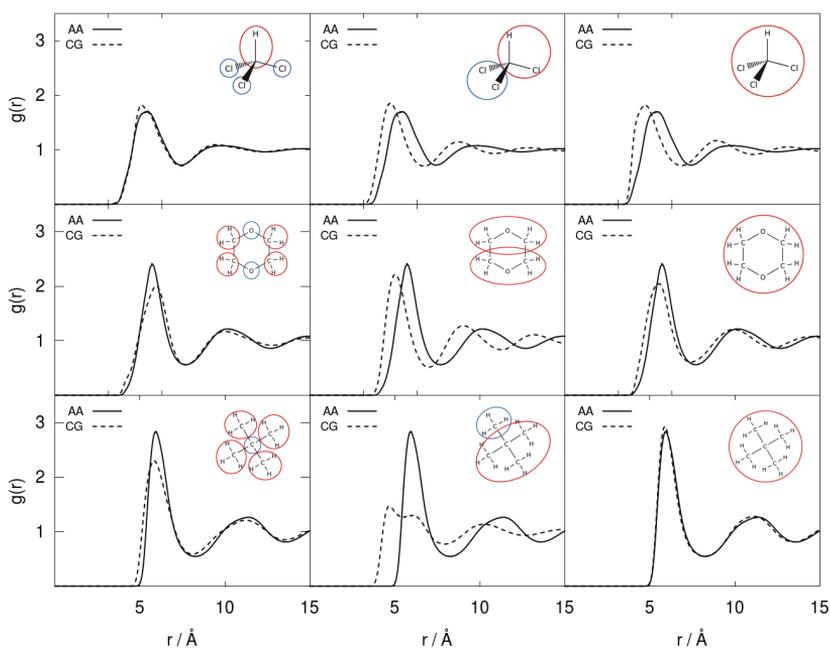


Figure 6.2: Radial distribution functions between the center of mass of the molecules computed from AA (continuous line) and CG trajectories (dashed line). To assist the eye the plots are labeled with a visual representation of the mapping scheme. From left to right, the com-RDF of chloroform, 1,4-dioxane and neopentane are shown.

Figure 6.2 compares the radial distribution functions of the center of mass of the CG models to their atomistic reference. In most cases the CG models reproduce the radial distribution functions of the underlying atomistic system, with some prominent exceptions. The single bead mapping is particularly successful in describing the RDF of neopentane, the united-atom coarse-graining is effective in capturing the center of mass RDF of atomistic chloroform. As far as the dumbbell of neopentane is concerned, the heterogeneity in the bead-size affects the quality of the model. Having a molecule with tetrahedral symmetry, such as neopentane, represented as an asymmetric two bead object leads to discrepancies in the RDFs. The first peak is split into two peaks. The CH<sub>3</sub> beads are able to come closer to each other (first peak), when compared to the much larger C<sub>4</sub>H<sub>9</sub> beads (second peak). The mismatch between AA and single bead chloroform radial distribution functions can be attributed to the approximation of describing an anisotropic molecule with a spherically symmetric potential. For example, a chloroform ( $C_{3v}$ ) molecule has lower symmetry with respect to tetrachloromethane ( $T_d$ ) and is further away from an isotropic sphere representation. That is the reason why, as shown in a previous work, the agreement between AA and CG is better in the case of tetrachloromethane.[256] The shift to the left in the radial distribution functions of some CG system suggests higher densities of the CG model. For comparison, in table 6.1 the average densities of the NPT simulations are reported. The CG systems generally have equal or higher density. As expected, the offset

Table 6.1: Densities of molecular liquids at different resolutions in units of g mL<sup>-1</sup>. The values are averages from NPT simulations (1 bar, 300 K (275 K for neopentane)) with AA and CG force fields.

level	chloroform	dioxane	neopentane
AA	0.853	0.596	0.390
UA	0.920	0.590	0.416
dumbbell	1.279	0.848	0.467
single bead	1.276	0.661	0.391

in the density is strong for the systems, where the RDF is shifted to smaller radii. In the dumbbell representation the liquids have the highest density, which coincides with this mapping scheme generally not being able to reproduce the structure of the reference.

## 6.4.2 Dynamics

In order to investigate the effects of coarse-graining on the dynamic modes of molecular liquids, we analyze three sets of trajectories: all-atom, mapped and coarse-grained. The purpose of the mapped trajectories is to establish reference systems that do not depend on any feature of the CG potentials. For each step of the atomistic trajectory we compute

---

---

where the location of the CG site would be and the respective velocity. We have mapped the AA trajectories to the three different degrees of coarse-graining. The mapping has been performed using the VOTCA package.[87] In fig. 6.3 we show the DoS of three different fluids, separated in translation, rotation and vibration. These represent the target dynamics that a CG system should ideally reproduce. The loss of some dynamic modes is inherent to the choice of the mapping. Therefore, we firstly discuss the changes in dynamics that stem from the choice of the mapping scheme, which is analyzed in the following.

The translational DoS, which is calculated from the velocity of the molecule's center of mass, shows typical liquid behavior for all three systems. It starts at a finite zero-frequency value, which is proportional to the diffusion coefficient, then reaches a maximum and decays to zero within  $200\text{ cm}^{-1}$ . Since the center of mass does not change with the mapping scheme, the translational DoS is unaffected.

The effect of mapping to the united-atom representation on the rotational DoF is small. This is expected, since the changes on the moment of inertia induced by a UA-mapping are negligible and the symmetry of the molecules remains unaffected. Mapping the molecules to a dumbbell representation introduces leaks of vibrational motion into the rotational motion, which manifest as peaks in the rotational spectrum. This is an artificial libration motion, which is not present in the atomistic trajectory. For example in the AA representation of neopentane there is a scissoring and rocking motions of the  $\text{CH}_3$  groups. When the molecule is mapped, parts of those vibrations become part of the rotation.

When the resolution of the model is lowered, the vibrational DoS function loses features. United-atom spectra have peaks from C–H stretching at around  $3000\text{ cm}^{-1}$  and bending motions of hydrogens at around  $1400\text{ cm}^{-1}$  removed. In the dumbbell vibrational spectra peaks are strongly reduced in intensity, because they only represent one DoF with  $\frac{1}{2}kT$  of kinetic energy. Those remaining peaks can be assigned to vibrational modes deforming the molecule's backbone structure. For neopentane in the united-atom spectrum two additional peaks at  $200\text{ cm}^{-1}$  and  $900\text{ cm}^{-1}$  disappear. We relate the peaks to the hindered rotation of the  $\text{CH}_3$  groups, which is lost in the UA mapping. For the single bead spectra, as expected, there is no vibration and rotation.

The dynamics of the mapped trajectories are interesting from the perspective of time scale separation, which is a recurring topic of interest in coarse-graining.[256, 264] Ideally, when interested in enforcing correct dynamics in CG systems with simple, memory-free friction terms, the mapping operator should separate out DoF that are faster and thereby adiabatically decoupled of the remaining ones. With this ideal mapping operator, the remaining DoF would behave Markovian, i.e. free of memory. By comparing the spectra of the AA system with the mapped trajectories, it is possible to get an idea of how fast (high frequencies) the lost dynamics and how slow (low frequencies) the remaining dynamics are, or to what extent they overlap. We see very good time scale separation for the UA mapping, where the neglected vibrational dynamics are much faster than the

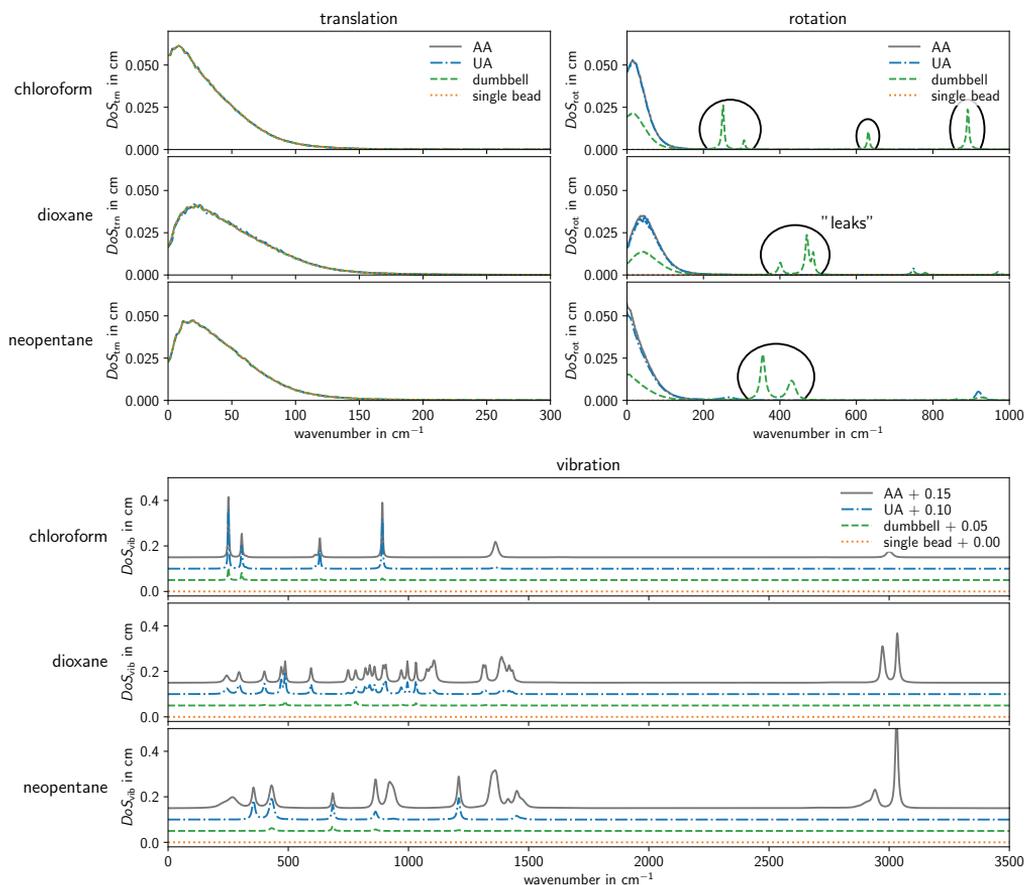


Figure 6.3: Translational (blue), rotational (red) and vibrational (black) DoS functions obtained from atomistic MD trajectories of three molecular liquids. The solid lines show the DoS functions of the AA liquid; the dashed lines are DoS functions of mapped trajectories. For better distinguishability, the vibrational DoS have been shifted along the y-axis. Leaks of vibrational motion into the rotational spectra have been highlighted.

---

---

remaining dynamics, with the exception for neopentane which was explained above. For the dumbbell mapping, the results indicate generally a bad separation. The remaining vibrational DoF are not clearly separated from the ones lost from the mapping. Mapping to a single bead at first view seems promising for a good time scale separation, since all vibrational DoF are neglected and are faster than the remaining translational dynamics. However, rotational DoF are also removed and those have, for the molecules considered in this study, peaks at the same frequency regime below  $100\text{ cm}^{-1}$  as the translation. In other words, orientational and translational motion happen on a similar time scale and do therefore likely couple. This depicts a poor time scale separability for molecules that are mapped to a single bead. We point out, that these results are related to small molecules and that for large molecules, where internal DoF dominate, we expect different behavior.

In fig. 6.4 we present the DoS functions of the CG liquids. For each level of coarse-graining the fluid has been simulated with potentials derived with the CRW method. With higher levels of coarse-graining the translational DoS functions of all molecules show a shift to lower frequencies. This is consistent with the general trend of having a smoother potential energy surface (PES) at higher levels of coarse-graining, which leads to faster diffusion and increased values at zero and low frequencies. For dioxane and neopentane the peaks of the CG systems are smaller and more intense at lower frequencies compared to the AA case. This change in dynamics has been described as shifts in the autocorrelation function of the molecular center of mass velocity.[223, 256] We interpret this trend as a side effect of the PES smoothing: with atomistic details missing the friction between two molecules moving against each other is lowered whereby the molecules oscillate more freely within the potential well. When approaching the extreme case where the PES is a single harmonic potential well, the DoS would consist of a single peak. Additionally, by coarse-graining to a sphere with isotropic potentials the directionality of the interaction is lost. This decreased variety in the intermolecular interaction leads to more uniform oscillations and more pronounced peaks in the spectra. There is one exception to the trend of higher diffusion coefficients with higher level of coarse-graining: the dumbbell neopentane has a higher intensity at low frequencies compared to the single bead representation, even though it has higher liquid density (compare table 6.1). This could be an effect of mapping neopentane to an asymmetric dumbbell, which also leads to a poor structural representation of the liquid as seen in section 6.4.1.

The united-atom CG model generally represents the dynamics of the underlying atomistic simulation accurately. However, in terms of rotational dynamics we observe a strong deviation from the atomistic reference. As the molecule rotates, its movement in the AA representation is hindered by the presence of hydrogen atoms, which can be seen as teeth on a gear. Upon coarse-graining the teeth are lost and the free rotation of the molecules becomes more pronounced, which unfolds in a shift to lower frequencies in the rotational spectrum. This change in rotational dynamics is small for chloroform, where a single hydrogen atom is coarse-grained.

By going from a three dimensional to a linear dumbbell molecule, the number of

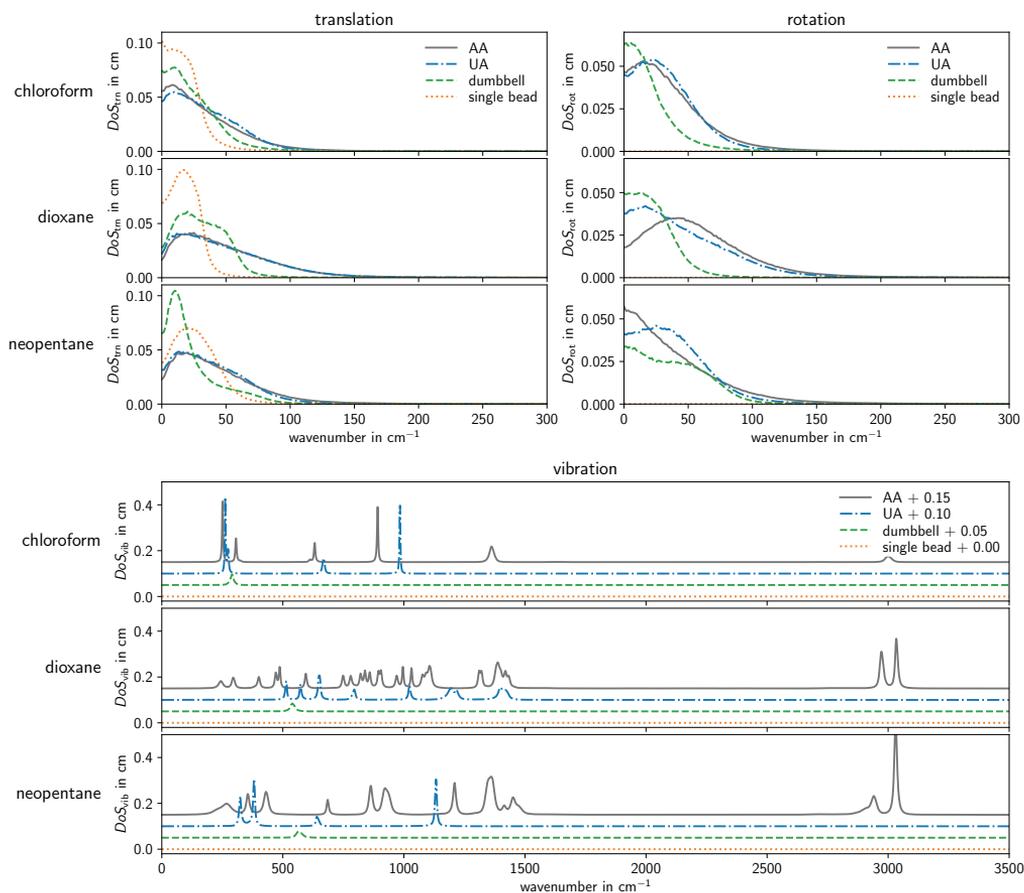


Figure 6.4: DoS functions obtained from the coarse-grained MD simulations are represented in different colors: translational (blue), rotational (red) and vibrational (black). The DoS functions of the systems of different levels of coarse-graining are presented in dashed lines, where for the vibrational DoS the lines have been shifted in the y-direction for clarity. As reference the AA DoS is included and shown as solid lines.

rotational DoF is reduced from three to two. This change is reflected in the integral of the DoS function. In addition to the loss in intensity we see a shift to lower frequencies, which shows that the dumbbell molecules rotates more freely. For chloroform and dioxane most of the spectrum is shifted to the region below  $50 \text{ cm}^{-1}$  with a high rotational diffusion coefficient. For neopentane the trend is less clear and at zero frequency the spectrum is lowered, contradicting the general trend of higher rotational diffusivity. The mapping of neopentane, a highly symmetric molecule, onto a linear asymmetric dumbbell has a negative impact on the representability of the dynamics, visible as well in the poor reproduction of the translational DoS as discussed earlier.

The vibrational DoS of the CG molecules in fig. 6.4 reproduce mostly the mapped trajectory counterparts, shown in fig. 6.3. Similar to the mapped spectra, hydrogen stretching and bending peaks are removed from the UA spectra. In contrast to the mapped DoS functions, there are generally less peaks and some of them are shifted. In the dumbbell representation vibrational DoS there is a single peak, since there is only one bond with one spring constant. These findings can be related to the Boltzmann inversion (BI) method, that is used to derive the potentials for the internal DoF. Since it is a structural method, BI aims to reproduce the distributions of the bonds, angles and dihedrals, but it does not necessarily lead to a reproduction of dynamics. However, internal modes are mostly harmonic and their dynamic behavior is well reproduced by BI. Not clear is the role of the peaks around  $300 \text{ cm}^{-1}$  in the vibrational spectrum of dioxane, that are missing in the UA representation.

### 6.4.3 Thermodynamics

In coarse-graining the change in entropy is of interest, since by merging atoms into single beads part of the information is irretrievably lost, which entropy is a measure for. An opposing contribution is caused by the softening of the potential energy surface, which enables the sampling of more microstates, resulting in an increase in the entropy of the system. All entropies, calculated from the DoS, are shown in fig. 6.5. Those entropies have been obtained by using eqs. (6.1) to (6.4). Note that we use the terminology of the 2PT authors[196], specifically translational, rotational and vibrational entropy, where positional, orientational and internal would possibly be more precise, since entropy is not a measure of a particular motion, but of the amount of phase space that is sampled by the motion.

The entropy  $S_R$ , calculated from the mapped trajectory, represents the CG phase space, that is accessible from the atomistic samples. The difference to the AA entropy  $S_r$ , is the mapping entropy

$$S_{\text{map}} = S_r - S_R. \quad (6.6)$$

$S_r$  is shown in fig. 6.5 as the AA value in either column of the plot,  $S_R$  is represented by the values in the first row. This mapping entropy is not identical to the (configurational)

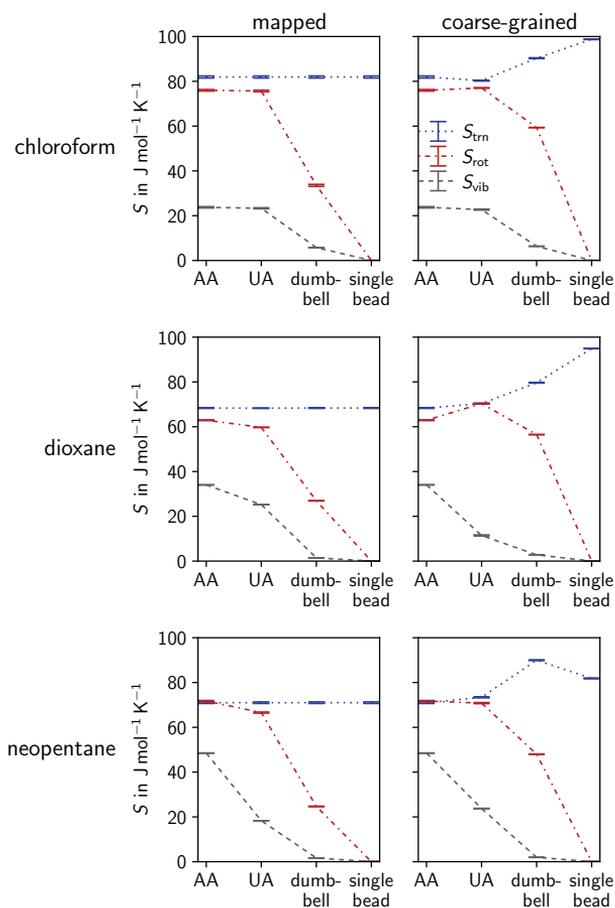


Figure 6.5: Entropy changes due to coarse-graining of three molecular liquids. The first column shows changes in translational, rotational and vibrational entropy from mapping the AA trajectory. In the second column the changes of those entropies upon coarse-graining are presented. The error bars result from the analysis of five consecutive blocks in each trajectory.

mapping entropy  $S_{\text{map,conf}}$  defined in the relative entropy framework[56, 236]. The difference is that the latter is expressed in terms of configurational excess entropies

$$S_{\text{map,conf}} = S_{r,\text{conf}}^e - S_{R,\text{conf}}^e. \quad (6.7)$$

Therefore, the two mapping entropies differ by the offset of ideal gas entropies of the mapped and atomistic system

$$S_{\text{map}} = S_{\text{map,conf}} + (S_r^{\text{id}} - S_R^{\text{id}}). \quad (6.8)$$

The ideal gas term is analytical but not trivial, since the composition of atoms and their masses changes through the mapping. It includes changes from integration of the momentum space, which has lower dimensionality in the mapped system. We expect, that the Gibbs inequality argument made by Rudzinski and Noid[236] still holds and that the mapping entropy is always non-negative. Because the coarsening lowers the dimensionality of the coordinate space, the configurational mapping entropy is always positive.[31] The same argument should hold for the momentum space, which is similarly reduced by the mapping operator. The monotonous decay of all entropies in fig. 6.5 (left panel) indicates that this holds not only for the total mapping entropy, but also for each component of the mapping entropy.

The argument of reduced phase space dimensionality through the mapping can be phrased more simply as the loss of DoF. It unfolds in a zero loss of translational DoF at all levels of resolution, in the loss of few vibrational DoF when effacing the single hydrogen of chloroform, in the loss of most vibrational DoF of neopentane in the dumbbell scheme and in the loss of all rotational and vibrational DoF in the case of all single bead representations. However, the linear relation

$$S_R \approx \frac{n_{\text{DoF}}(R)}{n_{\text{DoF}}(r)} S_r \quad (6.9)$$

does not apply, as for example the rotational entropy does drop by significantly more than  $\frac{1}{3}$  when going from AA to dumbbell. The mapping entropy needs a more careful assessment. The nonlinearity of the scaling has previously been observed in the case of the study of model protein domains.[31]

Since the center of mass of each molecule is preserved in the differently mapped representations, the translational entropy stays constant

$$S_{\text{map,trn}} = 0. \quad (6.10)$$

In the rotational contribution, we see a small negative mapping entropy for the united-atom level. It does not originate from changes in dynamics, but from changes of the molecules moment of inertia. The latter is generally lowered, when the position of the CG site coincides with the center of mass of its atoms. The dumbbell representations show a

drop of rotational entropy, that is larger than the expected 33 %. Instead the loss of entropy is about one half, because of the leaks in the rotational DoS, which effectively blueshift parts of the spectrum. The vibrational mapping entropy of united-atom chloroform is close to zero, because only high frequency modes are lost. In contrast, the vibrational mapping entropy is substantial for neopentane, where hydrogen is responsible for low frequency peaks.

The graphs in the second column of fig. 6.5 represent the entropy components of the CG molecules. It is worth noting, that it generally follows the trend of the mapping entropy, since it is mainly governed by the loss of DoF, which is the same in mapped and CG trajectories. Additionally, effects from the imperfections in the CG potential now come into play and are visible in the difference to the mapping entropy, which is shown in the first column. By increasing the level of coarse-graining the potential energy landscape is progressively smoothed out and the translational entropy is growing, with the exception of dumbbell neopentane, which stands out. The effect of coarse-graining to a UA level on the translational or rotational entropy is small. However, the vibrational entropy is noticeably reduced, and, similar to our findings with the mapping entropy, this is related to the unexpected loss of low frequency modes, that are caused by hydrogen atoms in the AA model. In the rotational DoS in fig. 6.4, we see a shift to low frequencies, which generally leads to higher entropy, especially for the dumbbell molecules. The CG molecules rotate more freely than their atomistic counterparts, thereby leading to higher rotational entropy compared to the mapped trajectory. Since the number of DoF is reduced by one-third, the rotational entropy is still lower than in the AA system. This shows how a reduction in phase space dimensionality and a smoothed potential energy surface can be contrary contributions.

The 2PT method infers the entropy from dynamic modes, which is a non-obvious connection. A similar relation, which was initially found by Rosenfeld, can be made using the excess-entropy scaling, which relates the diffusion coefficient of a fluid to its excess entropy [225]

$$D \propto \exp(\alpha S^e / k_B) \quad (6.11)$$

where  $\alpha$  is a numerical constant. This relationship is theoretically justified but has several exceptions.[265] When used as a predictor for entropies from dynamics it shows a certain similarity to the 2PT method, which starts from the (translational) DoS and therein the diffusion coefficient. In both methods a redshift towards diffusive modes is connected to a larger entropy. The theory of excess entropy scaling has been applied by Armstrong et al. to investigate the relation of atomistic and CG water.[266] They found the two representations each separately to fulfill the Rosenfeld scaling relation and that the difference comes from the loss of the rotational entropy. Shell makes systematical connections between the excess entropy scaling and the relative entropy methods.[267] He compares LJ-fluids with soft-sphere potentials derived with the relative entropy method and finds that for a low relative entropy the diffusivity is mostly conserved. In an attempt

to test the excess-entropy relation with our CG systems we use the (2PT) translational entropies to calculate translational excess entropies

$$S_{\text{trn}}^e = S_{\text{trn}} - S_{\text{trn}}^{\text{id}} \quad (6.12)$$

where  $S_{\text{trn}}^{\text{id}}$  is the ideal gas entropy from the Sackur-Tetrode equation.[268] The diffusion coefficient  $D$  we take from the zero frequency value of the translational DoS[195]

$$D = \frac{D_0 S_{\text{trn}}(0) k_B T}{12m} \quad (6.13)$$

where  $m$  is the mass of the molecule. fig. 6.6 shows, that the entropies calculated generally

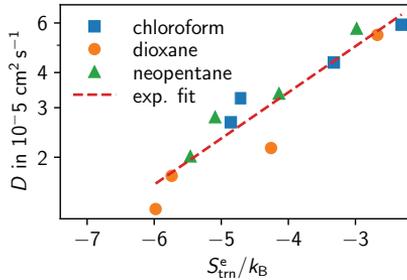


Figure 6.6: Diffusion constants and excess entropies of CG molecules at different levels of coarse-graining plotted with a logarithmic y-axis. The exponential fit has the form  $D = 15.3 \times 10^{-5} \text{ cm}^2 \text{ s}^{-1} \exp(0.381 S_{\text{trn}}^e / k_B)$ .

follow the excess-entropy relation but the correlation is not very strong. Also the constant  $\alpha$  is usually around 0.8 [265] where here we find it to be 0.381. The choice to exclude rotational and vibrational contributions at this point is justified by that the theory of excess-entropy scaling does not normally include any notion of those DoF. Indeed the relation is mostly tested with entropies calculated from the multiparticle-correlation expansion up to the second or third term using only the RDF and no orientational information.[269–271] It was found previously that for molecular systems the relation has different form for different molecular shapes.[272] This can be part of the explanation for why the data points are not perfectly fulfill the scaling relation since the different CG representations used in this work have different structures.

By summing the contributions of translation, rotation and vibration, the total entropy of the liquid is obtained and the values are given in table 6.2. A comparison with experimental values shows that the entropy of the fluids is underestimated, which Pascal et al. attribute to shortcomings of the force field.[240] For AA chloroform and 1,4-dioxane the entropy has been evaluated with the 2PT method before, in the work of Pascal et al., with the same force field and at similar state point.[240] We find that our entropies are

about 5 % lower and we attribute the difference to the rotational DoS, which we calculated with the method from [196], while in the reference paper  $DoS_{rot}$  equals  $DoS_{total} - DoS_{trn}$  (see figure 2d of Pascal et al.).

Improved force fields tuned specifically to reproduce molecular vibrations should lead to a better estimation of the entropy.[240] The total entropy of the liquids strongly decrease with higher levels of coarse-graining. The loss is about equal for mapped and CG systems at the UA level. At lower resolutions the decrease is more drastic for the entropies of the mapped trajectory. Here the CG systems have a higher entropy caused by their “softened” force field. The largest difference between mapped and CG system is found for the dumbbell representation. We can relate this discrepancy to the rotational entropy, which is amplified because the CG molecules rotate far more easily, as seen in section 6.4.2.

Table 6.2: Molar entropies calculated from the CG systems and mapped trajectories (R). The entropy of the fluid is the sum of the translational, rotational and vibrational contribution. Experimental values are taken from the NIST Database.[273] All values are in  $\text{JK}^{-1} \text{mol}^{-1}$ .

molecule	level	$S_R$	$S_{CG}$	$S_{exp}$	$S_{map}$	$(S_{CG} - S_R)$
chloroform	AA	181.6				
	UA	180.9	180.0		0.8	-0.9
	dumbbell	121.2	156.0		60.4	34.8
	single bead	81.9	98.8		99.7	16.8
dioxane	AA	165.3		196.6		
	UA	153.3	152.1		12.0	-1.2
	dumbbell	96.7	138.9		68.5	42.2
	single bead	68.3	94.9		96.9	26.5
neopentane	AA	191.2		216.2		
	UA	155.9	167.9		35.3	12.1
	dumbbell	97.3	139.9		93.9	42.6
	single bead	71.0	81.8		120.2	10.8

In summary, when computing the change in entropy in moving from an atomistic to a CG description of the system, two opposing contributions have to be considered. On one hand the loss in entropy due to a reduction of the DoF, on the other an increment in the translational and rotational entropic terms due to a flattening of the potential energy landscape in the CG system. With techniques as thermodynamic integration or test particle insertion it is hard to look at these effects, however with the 2PT model we are given the unique opportunity, even though approximative, to separate these contributions. The reason is that it derives the entropy from the dynamics, more specifically the DoS of the

---

---

system, which we can calculate for the CG systems, as well as for the mapped trajectories, which are free of effects from the CG potentials. Care should be taken in applying the 2PT approach to CG simulations in which friction terms are added, as for example with dissipative particle dynamics. While the dynamics of the system is dampened, the energy levels and the partition function of the system are unchanged, because they are set by the conservative interactions only. But with the 2PT model a change in the dynamics, even when sampling the same configurations, would result in a shift in entropy. This is a limitation of the 2PT model which assumes that the dynamics result purely from the interaction potentials. It implicitly requires a thermostat that changes the dynamics of the system only minimally and assumes no additional frictions. While inferring the entropy from the dynamics may support our understanding on the thermodynamics of CG models, it may not be applicable under all circumstances.

## 6.5 Conclusion

The change in entropy introduced by coarse-graining is model- and system-dependent. Entropy cannot be inferred a priori from the loss of the DoF alone and an additional way of computing the entropy variation is required.

In this work, the 2PT model is applied to the study of the vibrational, rotational and translational entropy of molecular liquids at different resolution. We separate the contribution to the entropy of the system that is coming from the choice of a mapping operator, which is referred to as mapping entropy, from the overall entropy of the system, which is affected by the imperfections of the CG potentials. Previous work by Baron et al. [231] showed how for systems that are dominated by intramolecular DoF, like lipid chains, the total configurational entropy coincides with the entropy computed from the mapped trajectory. Here, we observe that the standard molar entropy of CG molecular liquids differs from the mapping entropy. As a byproduct of computing entropy with the 2PT method, information concerning the dynamics in form of the vibrational density of states is collected. The analysis of the DoS helped us understand which vibrational modes are “washed out” by coarse-graining and how the dynamics of the system is affected. The CG models are able to mimic, for the most part, the features of the target DoS, which is computed from the rerun of the atomistic simulation. Minor leaks of the vibrations are found in the rotational spectra of the mapped trajectory. United-atom models reproduce translational dynamics effectively and rotational dynamics poorly. We observe a deviation between the DoS of the dumbbell representation of neopentane, which likely stems from the choice of a highly asymmetric mapping. The procedure presented in this article can indeed be used as a tool to screen different mappings, aiming at the optimal representation of the molecule. Importantly, the 2PT model can complement effectively other methods that estimate entropy, but are restricted in considering the configurational contribution or do not resolve the mappings in terms of spectra. In perspective, this approach can be

---

used to verify whether slow and fast DoF of a given system occur on different time-scales and can be clearly separated. A specific representation of the system may allow for a better time-scale separation of slow and fast modes, and lead to a better CG model.

## **Acknowledgments**

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 233630050 - TRR 146 on “Multiscale Simulation Methods for Soft Matter Systems”.

---

## 7 Conclusions and Outlook

---

In this work, the method development, application, and effects of coarse-graining molecular systems have been explored. In chapters 3 and 4, substantial progress in developing fast and stable iterative coarse-graining methods was presented. Integral equation (IE) theory has been successfully integrated into Newton's method to retain a target radial distribution function (RDF). The method provides a good approximation to the Jacobian and is comparable with the exact Jacobian inverse Monte Carlo (IMC) while significantly faster. For flexible molecules, the integral equation theory is found to give a less precise Jacobian, resulting in a convergence roughly half as fast as IMC. For both methods, IE-based coarse-graining and IMC, instabilities in the repulsive flank of pair potentials were understood. The problem was found to be an overshoot of Newton's method in the RDF onset region, where large relative differences in the value of the RDF occur. For that, a stabilizing modification to Newton's method was proposed. Also, a Gauss-Newton method is explored which works with the IE-based and the IMC Jacobian and which allows for constraining a potential to a thermodynamic property. It was shown how a constraint is derived for any thermodynamic function expressible in terms of pair potential, pair force, RDF, and pair distance. For example, this formulation enables a model for methanol-water mixtures with implicit water in the CG representation and pressure that matches the osmotic pressure of the atomistic reference. A combination of constraints is possible, for example, to develop a coarse-grained (CG) model of hexane which matches the reference's pressure and enthalpy of vaporization. However, for some combinations of constraints, the algorithm would not converge.

The methodological advancements presented make the application of structural coarse-graining methods faster and easier. Since Newton's method (with either the IE or IMC Jacobian) is now almost as stable as iterative Boltzmann inversion (IBI), they should be preferred, because they always converge to the same potential. In the future, the newly developed IE-based Jacobian could be used with other coarse-graining methods. Concretely, IMC has previously been found to be numerically equivalent to the relative entropy (RE) method applied to free-form pair potentials.[91] This equivalence implies that IE theory could be used to increase the speed of the RE method. Also, the findings and improvements in the stability of Newton's method might be transferable to the RE method. Besides, the Gauss-Newton method could be extended to approximately fit RDFs with parametrized potentials. The needed Jacobian elements for the potential parameter  $p_j$  could be obtained from the known Jacobian with respect to a tabulated potential ( $\frac{\partial g_i}{\partial u_i}$ )

---

---

by

$$\frac{\partial g_i}{\partial p_j} = \sum_l \frac{\partial g_i}{\partial u_l} \frac{\partial u_l}{\partial p_j}. \quad (7.1)$$

The derivative of the pair potential  $u_l$  at distance  $r_l$  with respect to parameter  $p_j$  should be obtainable numerically, e.g. by differentiating a fit function. While the Mie-potential is probably too simple for most CG models, it has the upside of simplicity and computational speed. However, more complex models, for example, with a term to allow for local maxima in the potential function, are thinkable. Packages such as OpenMM have started providing GPU support for arbitrary potential functions, thereby lowering the performance hurdle for non-standard potential forms. Moreover, a logical extension to the proposed Gauss-Newton method is a multistate formulation. Multistate IBI exists, but it involves a manual weighting of IBI updates at different state points or compositions.[129] Instead, with the Gauss-Newton method, it would be possible to minimize the squared distance between RDFs at multiple state points by optimizing a single pair potential. If this is meaningful will depend strongly on the difference between the state points and the kind of transferability that is required to capture them with a single potential. For example, CG models are generally not well temperature transferable, except if the pair potentials are explicitly temperature dependent.[150] Nevertheless, for mixtures at different concentrations, a set of pair potentials that represent the structure well across different compositions might be found.[274] A multistate Gauss-Newton method has been implemented in VOTCA and was planned to be part of chapter 4 but has not been thoroughly tested. The algorithmic goal is to optimize pair-potentials to yield the best possible match of the RDF at multiple states (index  $s$ ) which can have different thermodynamic conditions or different compositions. For a single component system, one would vary the pair potential to minimize

$$\left\| \sum_s g_s^{\text{CG}} - g_s^{\text{tgt}} \right\|_2, \quad (7.2)$$

but this can be straightforwardly generalized to multicomponent systems. In contrast to multistate IBI, this formulation has a clear objective. Since the implementation in VOTCA exists, this work could easily be continued and compared with multistate IBI.[129] Regarding the constraints, they allow much more flexibility and control when constructing CG models. It would be a compelling prospect to extend the work on combinations of constraints. There is currently no physical understanding of why it is impossible or very hard to have certain combinations of constraints, e.g., a single bead water model with pair potentials that represents pressure, compressibility, and RDF at the same time.

In chapter 5, iterative coarse-graining methods were utilized to parametrize ion-water pair potentials in atomistic force fields. No *real* coarse-graining, i.e., merging of atoms into beads, took place. The reference was ab initio molecular dynamics (AIMD), and the derived model was an existing atomistic model with only the ion-water interactions changed. One full charge model and one scaled charge model (electronic continuum cor-

---

---

rection, ECC) were modified to reproduce the ion-oxygen RDF from the AIMD calculation. The removed degrees of freedom are the electrons that are explicitly resolved at the AIMD level. The resulting potential was found to have a less steep repulsive flank than the  $r^{-12}$  term of the commonly used Lennard-Jones (LJ) potential. This effect was larger for ions with higher charge density ( $\text{Li}^+$ ,  $\text{Ca}^{2+}$ ) and smaller for ions with lower charge density ( $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Cl}^-$ ). While the repulsion between atoms is undoubtedly not well represented in most systems, the situation between cations and water molecules is particular. Due to the water molecule's high dipole moment and the cation's high charge density, there is a strong Coulomb attraction. This attraction gives the repulsive flank of the cation-oxygen pair potential much more sampling than it happens in systems with non-polar molecules. The rattling of the cation against its hydration shell is directly affected by the repulsive potential. With Lennard-Jones (LJ) potentials, the frequency was found systematically too high, while the AIMD-derived pair potentials matched experimental peak positions. The effects of the misrepresentation are not limited to dynamical properties. It was found that the AIMD-derived potentials improve the solvation entropies, which are systematically smaller than experimental values with LJ potentials. A different issue with the strong Coulomb attraction between cations and water molecules is that it dominates the attractive part of the LJ potential. The effect leads to a situation where only the repulsive part  $4\varepsilon(\frac{\sigma}{r})^{12}$  is important. Consequently, many cation LJ force fields have very different parameters  $\sigma$  and  $\varepsilon$ , but with similar values of  $\varepsilon\sigma^{12}$ .

Building on those findings, it seems that the LJ potential should probably be disestablished for the interactions between highly charged atoms. A faithful representation of the atom-atom repulsion can also be expected to play a principal role for non-polar molecules at very high pressures. High pressures are of interest in many fields such as geochemistry, deep-sea biochemistry, or engineering of polymerization processes.[275] A systematic study of the shape of the repulsive potential between atoms in different functional groups could lead to a better understanding of atom-atom interactions and their consequences. Iterative methods for structure matching offer a systematic approach to that end. The general automated atomic model parameterization (GAAMP) method does something similar, although it is fixed on the LJ potential.[276] If one is to follow the top-down parametrization route, the findings on the vibrational frequencies can be used. The frequencies of the rattling of the ion in the solvation shell have proved to be sensitive to the steepness of the repulsive potential. Thus, the systematic parametrization could use these frequencies as an additional target property and use potential forms such as Buckingham (exp-6) or Mie (m-n or m-6). A shortcoming of our ion-water potentials was that the ion-ion interactions were left untouched. This led, in the models with unscaled charges, to an imbalance in ion-ion and ion-water attractions and overly much ion pairing. It would be interesting to know whether this did not happen with scaled charges because those models are somehow "better". In general, both charge scaling and softer repulsion potentials seem to improve the representability of electrolyte force fields. How much the two are orthogonal or interdependent, future studies have to explore.

---

---

Back to ion-pairing: a possible way to improve the results with unscaled charges would leverage constraints on the potentials. Concretely, a constraint that fixes the ion-water potential energy, including the electrostatic interaction energy, is expected to preserve the balance between the different interactions. On the other hand, the interaction between cations and anions is probably also poorly described by an LJ potential as findings for molten salts show.[183] Hence, a bottom-up parametrization is desirable also for the ion-ion parameters. Those could, for example, be obtained with special sampling techniques in AIMD simulations.[277] It remains to see what effects the misrepresentation of ion-water potentials has in more complex systems such as proteins or polymers in electrolyte solutions. Existing LJ models often get the solvation free energy of the salts at infinite dilution right but have solvation entropies that are too negative. Accordingly, the solvation enthalpy is too low (too negative). For studies that investigate the effects of ions on macromolecules at different temperatures, this will lead to an imbalance of the energy and entropy term. But also at a constant temperature, an effect can be expected. The solvation shell, which is too structured with LJ models, and its properties are important when ions come into direct contact with other molecules because in the process, water molecules lose contact with the ions.

In chapter 6, the changes in the vibrational density of states (VDOS) and entropy upon changing the resolution in coarse-graining were assessed. To that end, the two-phase thermodynamic (2PT) model was used, which allows for a detailed (but approximate) look at the contributions of the molecular entropy. Three molecular liquids with four different resolutions: atomistic, united-atom, dumbbell, and single bead, were the subject of this investigation. It was shown that a large part of the entropy loss with decreasing resolution comes from the mapping alone. A significant factor is the loss of rotational degrees of freedom upon going to the dumbbell and single-bead mapping. However, the CG potential plays a role and compensates the entropy loss to some degree by an increase in translational entropy. The change in the dynamics was also investigated. It was found that united-atom models show good agreement with the atomistic resolution in the translational VDOS. Contrary, the rotational VDOS is much accelerated at the united-atom level due to the lost friction from the hydrogen atoms. In the future, a similar study could be performed on polymer systems, where CG models are of high usefulness. Different findings than for organic solvents are to be expected due to the higher number of intramolecular DOFs. Furthermore, CG models could be given artificial internal degrees of freedom that compensate for the entropy loss. The 2PT model offers a quick route to determine the entropy difference between the resolutions.

Towards the end of this thesis, a more speculative outlook into the future of molecular simulation and coarse-graining shall be taken. Moore's law is likely to continue and will increase the amount of available computational resources steadily. Larger and longer molecular dynamics (MD) simulations will be possible, which are out of reach today. However, that does not make CG models obsolete since there will always be interest in length scales that are not reachable by fine-grained models. Also, due to the many orders

---

of magnitude length and time scales needed for certain systems, CG representations will be needed for many years to come. Quantum computers are becoming increasingly better and might at some point accelerate quantum calculations significantly. The consequences of that are hard to predict, however, but it seems unlikely that AIMD will ever be faster than classical MD. With the current rise in machine learning, neural networks have found many applications in molecular simulations. Particularly interesting is the prospect of very efficient “force fields” with precision comparable to AIMD.[278] Yet, it is unlikely that classical force fields will be fully replaced, if only because they are much better understood and have a straightforward physical interpretation. Adaptive resolution schemes are an exciting advancement in coarse-graining. In current methods, the boundary between the schemes is always fixed.[279] But one could imagine a dynamic change of the boundary between differently resolved schemes. For example, a piece of polymer could be simulated with the finite element method, and, upon the application of stress deform. At some spatial point, the stress crosses a threshold, and a CG model connected to the surrounding finite element grid is introduced. The introduction of finer resolution domains, repeats at lower levels until there is some model that allows for the breakage of chemical bonds. For the construction of such a flexible cascade of models, many physical and algorithmic challenges would have to be faced. But ultimately it might provide profound insights on systems that scale many length and time scales. In the field of biochemistry, CG models will likely continue to be useful. Interestingly, the particular problem of protein folding was not solved with CG models, but by the data-driven machine learning model AlphaFold.[280] It is likely that machine learning, or a combination of the same with MD simulations, will continue to solve problems that are hard to come by with other methods. This is especially the case for problems, where a large amount of data exists (protein folding is one of those). Still, there are many problems regarding the function and arrangement of proteins, membranes, and parts of the cell, that will be best addressed using CG models.



---

# Reproducibility

---

The replication of experimental results is an important part of the scientific method. Computer simulations are no exception and need to be documented well in publications. Additionally, making simulation files, force field details, results, scripts, and software publicly available can greatly help to enable replication and reuse.

Under the following links one can find Jupyter notebooks that were central in the set-up and analysis of the computer simulations. Also included are simulation files and intermediate results that represent the data presented in the publications.

Paper 1 / Chapter 3

<https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2662>



Paper 2 / Chapter 4

<https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/3492>



Paper 3 / Chapter 5

<https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/3493>



Paper 4 / Chapter 6

<https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2788>





---

## Bibliography

---

- [1] K. Binder, *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, Oxford University Press, **1995**, 602 pp.
- [2] M. Karplus, J. A. McCammon, “Molecular Dynamics Simulations of Biomolecules”, *Nature Structural Biology* **2002**, *9*, 646–652.
- [3] S. A. Hollingsworth, R. O. Dror, “Molecular Dynamics Simulation for All”, *Neuron* **2018**, *99*, 1129–1143.
- [4] O. M. H. Salo-Ahen, I. Alanko, R. Bhadane, A. M. J. J. Bonvin, R. V. Honorato, S. Hossain, A. H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A. S. Larsen, E. Lescrinier, P. Marimuthu, M. U. Mirza, G. Mustafa, A. Nunes-Alves, T. Pantsar, A. Saadabadi, K. Singaravelu, M. Vanmeert, “Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development”, *Processes* **2021**, *9*, 71.
- [5] M. Feig, Y. Sugita, “Whole-Cell Models and Simulations in Molecular Detail”, *Annual review of cell and developmental biology* **2019**, *35*, 191–211.
- [6] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, “CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations”, *Journal of Computational Chemistry* **1983**, *4*, 187–217.
- [7] W. L. Jorgensen, J. D. Madura, C. J. Swenson, “Optimized Intermolecular Potential Functions for Liquid Hydrocarbons”, *Journal of the American Chemical Society* **1984**, *106*, 6638–6646.
- [8] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”, *Journal of the American Chemical Society* **1995**, *117*, 5179–5197.
- [9] X. Daura, A. E. Mark, W. F. Van Gunsteren, “Parametrization of aliphatic CH<sub>n</sub> united atoms of GROMOS96 force field”, *Journal of Computational Chemistry* **1998**, *19*, 535–547.
- [10] R. P. Feynman, “Simulating Physics with Computers”, *International Journal of Theoretical Physics* **1982**, *21*, 467–488.
- [11] P. Cieplak, F.-Y. Dupradeau, Y. Duan, J. Wang, “Polarization Effects in Molecular Mechanical Force Fields”, *Journal of Physics: Condensed Matter* **2009**, *21*, 333102.
- [12] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, “ReaxFF: A Reactive Force Field for Hydrocarbons”, *The Journal of Physical Chemistry A* **2001**, *105*, 9396–9409.

- 
- [13] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, "Comparison of Simple Potential Functions for Simulating Liquid Water", *The Journal of Chemical Physics* **1983**, *79*, 926–935.
- [14] M. W. Mahoney, W. L. Jorgensen, "A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions", *The Journal of Chemical Physics* **2000**, *112*, 8910–8922.
- [15] S. P. Kadaoluwa Pathirannahalage, N. Meftahi, A. Elbourne, A. C. G. Weiss, C. F. McConville, A. Padua, D. A. Winkler, M. Costa Gomes, T. L. Greaves, T. C. Le, Q. A. Besford, A. J. Christofferson, "Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations", *Journal of Chemical Information and Modeling* **2021**, *61*, 4521–4536.
- [16] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, "The Missing Term in Effective Pair Potentials", *The Journal of Physical Chemistry* **1987**, *91*, 6269–6271.
- [17] I. V. Leontyev, A. A. Stuchebrukhov, "Electronic Continuum Model for Molecular Dynamics Simulations", *The Journal of Chemical Physics* **2009**, *130*, 085102.
- [18] M. Kohagen, P. E. Mason, P. Jungwirth, "Accurate Description of Calcium Solvation in Concentrated Aqueous Solutions", *The Journal of Physical Chemistry B* **2014**, *118*, 7902–7909.
- [19] E. E. Bruce, N. F. A. van der Vegt, "Does an Electronic Continuum Correction Improve Effective Short-Range Ion-Ion Interactions in Aqueous Solution?", *The Journal of Chemical Physics* **2018**, *148*, 222816.
- [20] B. J. Kirby, P. Jungwirth, "Charge Scaling Manifesto: A Way of Reconciling the Inherently Macroscopic and Microscopic Natures of Molecular Simulations", *The Journal of Physical Chemistry Letters* **2019**, *10*, 7531–7536.
- [21] D. Horinek, S. I. Mamatkulov, R. R. Netz, "Rational Design of Ion Force Fields Based on Thermodynamic Solvation Properties", *The Journal of Chemical Physics* **2009**, *130*, 124507.
- [22] I. M. Zeron, J. L. F. Abascal, C. Vega, "A Force Field of Li<sup>+</sup>, Na<sup>+</sup>, K<sup>+</sup>, Mg<sup>2+</sup>, Ca<sup>2+</sup>, Cl<sup>-</sup>, and SO<sub>4</sub><sup>2-</sup> in Aqueous Solution Based on the TIP4P/2005 Water Model and Scaled Charges for the Ions", *The Journal of Chemical Physics* **2019**, *151*, 134504.
- [23] M. Fyta, R. R. Netz, "Ionic Force Field Optimization Based on Single-Ion and Ion-Pair Solvation Properties: Going beyond Standard Mixing Rules", *The Journal of Chemical Physics* **2012**, *136*, 124103.
- [24] G. Tóth, "Ab Initio Pair Potential Parameter Set for the Interaction of a Rigid and a Flexible Water Model and the Complete Series of the Halides and Alkali Cations", *The Journal of Chemical Physics* **1996**, *105*, 5518–5524.
- [25] A. P. Lyubartsev, K. Laasonen, A. Laaksonen, "Hydration of Li<sup>+</sup> Ion. An Ab Initio Molecular Dynamics Simulation", *The Journal of Chemical Physics* **2001**, *114*, 3120–3126.
- [26] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuzmaul, B. W. Lampson, D. Sanchez, T. B. Schardl, "There's Plenty of Room at the Top: What Will Drive Computer Performance after Moore's Law?", *Science* **2020**, *368*, eaam9744.

- 
- 
- [27] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, N. F. A. van der Vegt, “Systematic Coarse-Graining Methods for Soft Matter Simulations – a Review”, *Soft Matter* **2013**, *9*, 2108–2119.
- [28] L. G. Dunfield, A. W. Burgess, H. A. Scheraga, “Energy Parameters in Polypeptides. 8. Empirical Potential Energy Algorithm for the Conformational Analysis of Large Molecules”, *The Journal of Physical Chemistry* **1978**, *82*, 2609–2616.
- [29] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, O. Hahn, “Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates”, *Acta Polymerica* **1998**, *49*, 61–74.
- [30] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfandner, G. A. Voth, “A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules”, *Biophysical Journal* **2008**, *95*, 5073–5083.
- [31] T. T. Foley, M. S. Shell, W. G. Noid, “The Impact of Resolution upon Entropy and Information in Coarse-Grained Models”, *The Journal of Chemical Physics* **2015**, *143*, 243104.
- [32] M. Dallavalle, N. F. A. van der Vegt, “Evaluation of Mapping Schemes for Systematic Coarse Graining of Higher Alkanes”, *Physical Chemistry Chemical Physics* **2017**, *19*, 23034–23042.
- [33] M. Chakraborty, J. Xu, A. D. White, “Is Preservation of Symmetry Necessary for Coarse-Graining?”, *Physical Chemistry Chemical Physics* **2020**, *22*, 14998–15005.
- [34] D. Reith, B. Müller, F. Müller-Plathe, S. Wiegand, “How Does the Chain Extension of Poly (Acrylic Acid) Scale in Aqueous Solution? A Combined Study with Light Scattering and Computer Simulation”, *The Journal of Chemical Physics* **2002**, *116*, 9100–9106.
- [35] L. Ruiz, W. Xia, Z. Meng, S. Keten, “A Coarse-Grained Model for the Mechanical Behavior of Multi-Layer Graphene”, *Carbon* **2015**, *82*, 103–115.
- [36] S. Kloth, M. P. Bernhardt, N. F. A. van der Vegt, M. Vogel, “Coarse-Grained Model of a Nanoscale-Segregated Ionic Liquid for Simulations of Low-Temperature Structure and Dynamics”, *Journal of Physics: Condensed Matter* **2021**, *33*, 204002.
- [37] X. Periole, A. M. Knepp, T. P. Sakmar, S. J. Marrink, T. Huber, “Structural Determinants of the Supramolecular Organization of G Protein-Coupled Receptors in Bilayers”, *Journal of the American Chemical Society* **2012**, *134*, 10959–10965.
- [38] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, S. J. Marrink, “The Power of Coarse Graining in Biomolecular Simulations”, *Wiley Interdisciplinary Reviews. Computational Molecular Science* **2014**, *4*, 225–248.
- [39] S. Y. Joshi, S. A. Deshmukh, “A Review of Advancements in Coarse-Grained Molecular Dynamics Simulations”, *Molecular Simulation* **2021**, *47*, 786–803.
- [40] F. Müller-Plathe, “Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back”, *ChemPhysChem* **2002**, *3*, 754–769.
- [41] S. J. Marrink, A. H. de Vries, A. E. Mark, “Coarse Grained Model for Semiquantitative Lipid Simulations”, *The Journal of Physical Chemistry B* **2004**, *108*, 750–760.

- 
- [42] P. C. T. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. H. Bruininks, T. A. Wassenaar, P. C. Kroon, J. Melcr, V. Nieto, V. Corradi, H. M. Khan, J. Domański, M. Javanainen, H. Martinez-Seara, N. Reuter, R. B. Best, I. Vattulainen, L. Monticelli, X. Periole, D. P. Tieleman, A. H. de Vries, S. J. Marrink, “Martini 3: A General Purpose Force Field for Coarse-Grained Molecular Dynamics”, *Nature Methods* **2021**, *18*, 382–388.
- [43] C. Peter, K. Kremer, “Multiscale Simulation of Soft Matter Systems – from the Atomistic to the Coarse-Grained Level and Back”, *Soft Matter* **2009**, *5*, 4357–4366.
- [44] F. Ercolessi, J. B. Adams, “Interatomic Potentials from First-Principles Calculations: The Force-Matching Method”, *Europhysics Letters (EPL)* **1994**, *26*, 583–588.
- [45] S. Izvekov, G. A. Voth, “Multiscale Coarse Graining of Liquid-State Systems”, *The Journal of Chemical Physics* **2005**, *123*, 134105.
- [46] Y. Wang, W. G. Noid, P. Liu, G. A. Voth, “Effective Force Coarse-Graining”, *Physical Chemistry Chemical Physics* **2009**, *11*, 2002–2015.
- [47] E. Brini, V. Marcon, N. F. A. van der Vegt, “Conditional Reversible Work Method for Molecular Coarse Graining Applications”, *Physical Chemistry Chemical Physics* **2011**, *13*, 10468–10474.
- [48] R. L. Henderson, “A Uniqueness Theorem for Fluid Pair Correlation Functions”, *Physics Letters A* **1974**, *49*, 197–198.
- [49] F. Frommer, M. Hanke, S. Jansen, “A Note on the Uniqueness Result for the Inverse Henderson Problem”, *Journal of Mathematical Physics* **2019**, *60*, 093303.
- [50] A. P. Lyubartsev, A. Laaksonen, “Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach”, *Physical Review E* **1995**, *52*, 3730–3737.
- [51] A. K. Soper, “Empirical Potential Monte Carlo Simulation of Fluid Structure”, *Chemical Physics* **1996**, *202*, 295–306.
- [52] D. Reith, M. Pütz, F. Müller-Plathe, “Deriving Effective Mesoscale Potentials from Atomistic Simulations”, *Journal of Computational Chemistry* **2003**, *24*, 1624–1636.
- [53] S. Jain, S. Garde, S. K. Kumar, “Do Inverse Monte Carlo Algorithms Yield Thermodynamically Consistent Interaction Potentials?”, *Industrial & Engineering Chemistry Research* **2006**, *45*, 5614–5618.
- [54] D. Rosenberger, M. Hanke, N. F. A. van der Vegt, “Comparison of Iterative Inverse Coarse-Graining Methods”, *The European Physical Journal Special Topics* **2016**, *225*, 1323–1345.
- [55] M. Hanke, “Well-Posedness of the Iterative Boltzmann Inversion”, *Journal of Statistical Physics* **2018**, *170*, 536–553.
- [56] M. S. Shell, “The Relative Entropy Is Fundamental to Multiscale and Inverse Thermodynamic Problems”, *The Journal of Chemical Physics* **2008**, *129*, 144108.
- [57] A. Chaimovich, M. S. Shell, “Coarse-Graining Errors and Numerical Optimization Using a Relative Entropy Framework”, *The Journal of Chemical Physics* **2011**, *134*, 094112.

- 
- [58] V. Klippenstein, M. Tripathy, G. Jung, F. Schmid, N. F. A. van der Vegt, "Introducing Memory in Coarse-Grained Molecular Simulations", *The Journal of Physical Chemistry B* **2021**, *125*, 4931–4954.
- [59] P. K. Depa, J. K. Maranas, "Speed up of Dynamic Observables in Coarse-Grained Molecular-Dynamics Simulations of Unentangled Polymers", *The Journal of Chemical Physics* **2005**, *123*, 094901.
- [60] M. K. Meinel, F. Müller-Plathe, "Loss of Molecular Roughness upon Coarse-Graining Predicts the Artificially Accelerated Mobility of Coarse-Grained Molecular Simulation Models", *Journal of Chemical Theory and Computation* **2020**, *16*, 1411–1419.
- [61] H. Wang, C. Junghans, K. Kremer, "Comparative Atomistic and Coarse-Grained Study of Water: What Do We Lose by Coarse-Graining?", *The European Physical Journal E* **2009**, *28*, 221–229.
- [62] P. Ganguly, N. F. A. van der Vegt, "Representability and Transferability of Kirkwood-Buff Iterative Boltzmann Inversion Models for Multicomponent Aqueous Systems", *Journal of Chemical Theory and Computation* **2013**, *9*, 5247–5256.
- [63] N. J. H. Dunn, T. T. Foley, W. G. Noid, "Van Der Waals Perspective on Coarse-Graining: Progress toward Solving Representability and Transferability Problems", *Accounts of Chemical Research* **2016**, *49*, 2832–2840.
- [64] A. Das, H. C. Andersen, "The Multiscale Coarse-Graining Method. V. Isothermal-isobaric Ensemble", *The Journal of Chemical Physics* **2010**, *132*, 164106.
- [65] M. E. Johnson, T. Head-Gordon, A. A. Louis, "Representability Problems for Coarse-Grained Water Potentials", *The Journal of Chemical Physics* **2007**, *126*, 144509.
- [66] H.-J. Qian, P. Carbone, X. Chen, H. A. Karimi-Varzaneh, C. C. Liew, F. Müller-Plathe, "Temperature-Transferable Coarse-Grained Potentials for Ethylbenzene, Polystyrene, and Their Mixtures", *Macromolecules* **2008**, *41*, 9919–9929.
- [67] D. Rosenberger, N. F. A. van der Vegt, "Addressing the Temperature Transferability of Structure Based Coarse Graining Models", *Physical Chemistry Chemical Physics* **2018**, *20*, 6617–6628.
- [68] A. Villa, C. Peter, N. F. A. van der Vegt, "Transferability of Nonbonded Interaction Potentials for Coarse-Grained Simulations: Benzene in Water", *Journal of Chemical Theory and Computation* **2010**, *6*, 2434–2444.
- [69] E. J. Sambriski, G. Yatsenko, M. A. Nemirowskaya, M. G. Guenza, "Analytical Coarse-Grained Description for Polymer Melts", *The Journal of Chemical Physics* **2006**, *125*, 234902.
- [70] M. G. Guenza, M. Dinpajoo, J. McCarty, I. Y. Lyubimov, "Accuracy, Transferability, and Efficiency of Coarse-Grained Models of Molecular Liquids", *The Journal of Physical Chemistry B* **2018**, *122*, 10257–10278.
- [71] Q. Wang, D. J. Keffer, D. M. Nicholson, J. B. Thomas, "Use of the Ornstein-Zernike Percus-Yevick Equation to Extract Interaction Potentials from Pair Correlation Functions", *Physical Review E* **2010**, *81*, 061204.

- 
- [72] Q. Wang, D. J. Keffer, D. M. Nicholson, “A Coarse-Grained Model for Polyethylene Glycol Polymer”, *The Journal of Chemical Physics* **2011**, *135*, 214903.
- [73] S. Y. Mashayak, L. Miao, N. R. Aluru, “Integral Equation Theory Based Direct and Accelerated Systematic Coarse-Graining Approaches”, *The Journal of Chemical Physics* **2018**, *148*, 214105.
- [74] D. Levesque, J. J. Weis, L. Reatto, “Pair Interaction from Structural Data for Dense Classical Liquids”, *Physical Review Letters* **1985**, *54*, 451–454.
- [75] F. Delbary, M. Hanke, D. Ivanizki, “A Generalized Newton Iteration for Computing the Solution of the Inverse Henderson Problem”, *Inverse Problems in Science and Engineering* **2020**, *28*, 1166–1190.
- [76] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Reprint, Clarendon Press, Oxford, **1989**, 408 pp.
- [77] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, OUP Oxford, **2010**, 719 pp.
- [78] W. Gander, M. J. Gander, F. Kwok, *Scientific Computing - An Introduction Using Maple and MATLAB*, Springer International Publishing, **2014**.
- [79] K. R. Hadley, C. McCabe, “On the Investigation of Coarse-Grained Models for Water: Balancing Computational Efficiency and the Retention of Structural Properties”, *The Journal of Physical Chemistry B* **2010**, *114*, 4590–4599.
- [80] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, H. C. Andersen, “The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models”, *The Journal of Chemical Physics* **2008**, *128*, 244114.
- [81] C. N. Likos, “Effective Interactions in Soft Condensed Matter Physics”, *Physics Reports* **2001**, *348*, 267–439.
- [82] W. G. Noid, “Perspective: Coarse-grained Models for Biomolecular Systems”, *The Journal of Chemical Physics* **2013**, *139*, 090901.
- [83] S. T. John, G. Csányi, “Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials”, *The Journal of Physical Chemistry B* **2017**, *121*, 10934–10949.
- [84] M. S. Shell in *Advances in Chemical Physics*, (Eds.: Stuart A. Rice, Aaron R. Dinner), John Wiley & Sons, Ltd, **2016**, pp. 395–441.
- [85] C. Chen, C. Arntsen, G. A. Voth, “Development of Reactive Force Fields Using Ab Initio Molecular Dynamics Simulation Minimally Biased to Experimental Data”, *The Journal of Chemical Physics* **2017**, *147*, 161719.
- [86] T. Sanyal, M. S. Shell, “Transferable Coarse-Grained Models of Liquid–Liquid Equilibrium Using Local Density Potentials Optimized with the Relative Entropy”, *The Journal of Physical Chemistry B* **2018**, *122*, 5678–5693.
- [87] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, D. Andrienko, “Versatile Object-Oriented Toolkit for Coarse-Graining Applications”, *Journal of Chemical Theory and Computation* **2009**, *5*, 3211–3223.

- 
- 
- [88] V. Rühle, C. Junghans, “Hybrid Approaches to Coarse-Graining Using the VOTCA Package: Liquid Hexane”, *Macromolecular Theory and Simulations* **2011**, *20*, 472–477.
- [89] C. Jang, C. Abrams, “Thermal and Mechanical Properties of Thermosetting Polymers Using Coarse-Grained Simulation”, *The European Physical Journal Special Topics* **2016**, *225*, 1775–1783.
- [90] G. G. Rondina, M. C. Böhm, F. Müller-Plathe, “Predicting the Mobility Increase of Coarse-Grained Polymer Models from Excess Entropy Differences”, *Journal of Chemical Theory and Computation* **2020**, 1431–1447.
- [91] D. Rosenberger, T. Sanyal, M. S. Shell, N. F. A. van der Vegt, “Transferability of Local Density-Assisted Implicit Solvation Models for Homogeneous Fluid Mixtures”, *Journal of Chemical Theory and Computation* **2019**, *15*, 2881–2895.
- [92] F. Frommer, M. Hanke, A Variational Framework for the Inverse Henderson Problem of Statistical Mechanics, arXiv:2204.09293, **2022**.
- [93] L. S. Ornstein, F. Zernike in Proceedings of the Royal Academy of Sciences at Amsterdam, Vol. 17, **1914**, p. 793.
- [94] J.-P. Hansen, I. R. McDonald, *Theory of Simple Liquids*, Elsevier, **1990**, 569 pp.
- [95] T. Morita, “Theory of Classical Fluids: Hyper-Netted Chain Approximation, I: Formulation for a One-Component System”, *Progress of Theoretical Physics* **1958**, *20*, 920–938.
- [96] J. K. Percus, G. J. Yevick, “Analysis of Classical Statistical Mechanics by Means of Collective Coordinates”, *Physical Review* **1958**, *110*, 1–13.
- [97] J. L. Lebowitz, J. K. Percus, “Mean Spherical Model for Lattice Gases with Extended Hard Cores and Continuum Fluids”, *Physical Review* **1966**, *144*, 251–258.
- [98] F. Lado, “Pressure-Consistent Integral Equation for Classical Fluids: Hard-Sphere Solutions”, *The Journal of Chemical Physics* **1967**, *47*, 4828–4833.
- [99] J. A. Barker, D. Henderson, “Perturbation Theory and Equation of State for Fluids: The Square-Well Potential”, *The Journal of Chemical Physics* **1967**, *47*, 2856–2861.
- [100] A. A. Broyles, “Solutions to the Percus-Yevick Equation”, *The Journal of Chemical Physics* **1961**, *35*, 493–496.
- [101] A. Rotenberg, “Monte Carlo Equation of State for Hard Spheres in an Attractive Square Well”, *The Journal of Chemical Physics* **1965**, *43*, 1198–1201.
- [102] M. S. Wertheim, “Exact Solution of the Percus-Yevick Integral Equation for Hard Spheres”, *Physical Review Letters* **1963**, *10*, 321–323.
- [103] E. Thiele, “Equation of State for Hard Spheres”, *The Journal of Chemical Physics* **1963**, *39*, 474–479.
- [104] J. L. Lebowitz, “Exact Solution of Generalized Percus-Yevick Equation for a Mixture of Hard Spheres”, *Physical Review* **1964**, *133*, A895–A899.
- [105] W. R. Smith, D. Henderson, Y. Tago, “Mean Spherical Approximation and Optimized Cluster Theory for the Square-well Fluid”, *The Journal of Chemical Physics* **1977**, *67*, 5308–5316.

- 
- 
- [106] D. Chandler, H. C. Andersen, "Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids", *The Journal of Chemical Physics* **1972**, *57*, 1930–1937.
- [107] W. J. Briels, "Theory of Polymer Dynamics", *Lecture Notes; Uppsala University: Uppsala Sweden* **1994**.
- [108] C. M. Cortis, P. J. Rossky, R. A. Friesner, "A Three-Dimensional Reduction of the Ornstein–Zernicke Equation for Molecular Liquids", *The Journal of Chemical Physics* **1997**, *107*, 6400–6414.
- [109] A. Kovalenko, F. Hirata, "Hydration Free Energy of Hydrophobic Solutes Studied by a Reference Interaction Site Model with a Repulsive Bridge Correction and a Thermodynamic Perturbation Method", *The Journal of Chemical Physics* **2000**, *113*, 2793–2805.
- [110] T. Imai, A. Kovalenko, F. Hirata, "Hydration Structure, Thermodynamics, and Functions of Protein Studied by the 3D-RISM Theory", *Molecular Simulation* **2006**, *32*, 817–824.
- [111] D. P. Craig, I. R. Esum, "Chiral Discrimination in the Liquid State a Hard-Core Model for the Role of Repulsive Forces", *Chemical Physics* **1982**, *73*, 349–362.
- [112] D. Chandler, M. Rigby, P. A. Monson, D. J. Tildesley, W. B. Strett, W. R. Smith, P. A. Madden, J. M. Haile, W. A. Steele, J. S. Rowlinson, J. C. Dore, I. P. Gibson, J. L. Finney, G. Pálinkás, J. L. Rivail, H. J. C. Berendsen, P. Bordewijk, K. Singer, D. J. Adams, J. N. Murrell, E. B. Smith, "General Discussion", *Faraday Discussions of the Chemical Society* **1978**, *66*, 71–94.
- [113] D. Chandler, R. Silbey, B. M. Ladanyi, "New and Proper Integral Equations for Site-Site Equilibrium Correlations in Molecular Fluids", *Molecular Physics* **1982**, *46*, 1335–1345.
- [114] K. S. Schweizer, J. G. Curro, "Integral-Equation Theory of the Structure of Polymer Melts", *Physical Review Letters* **1987**, *58*, 246–249.
- [115] K. S. Schweizer, "Analytic RISM Theory of Polymer Alloys: Molecular Closure Predictions for Structurally Symmetric Blends", *Macromolecules* **1993**, *26*, 6033–6049.
- [116] F. Hirata, R. M. Levy, "A New RISM Integral Equation for Solvated Polymers", *Chemical Physics Letters* **1987**, *136*, 267–273.
- [117] A. Lyubartsev, A. Mirzoev, L. Chen, A. Laaksonen, "Systematic Coarse-Graining of Molecular Models by the Newton Inversion Method", *Faraday Discussions* **2010**, *144*, 43–56.
- [118] Y.-L. Wang, A. Lyubartsev, Z.-Y. Lu, A. Laaksonen, "Multiscale Coarse-Grained Simulations of Ionic Liquids : Comparison of Three Approaches to Derive Effective Potentials", *Physical Chemistry Chemical Physics* **2013**, *15*, 7701–7712.
- [119] B. L. Peters, K. M. Salerno, A. Agrawal, D. Perahia, G. S. Grest, "Coarse-Grained Modeling of Polyethylene Melts: Effect on Dynamics", *Journal of Chemical Theory and Computation* **2017**, *13*, 2890–2896.
- [120] A. Gooneie, S. Schuschnigg, C. Holzer, "A Review of Multiscale Computational Methods in Polymeric Materials", *Polymers* **2017**, *9*, 16.
- [121] A. P. Lyubartsev, A. Naômé, D. P. Vercauteren, A. Laaksonen, "Systematic Hierarchical Coarse-Graining with the Inverse Monte Carlo Method", *The Journal of Chemical Physics* **2015**, *143*, 243120.

- 
- 
- [122] C. D. Williams, M. Lísal, “Coarse Grained Models of Graphene and Graphene Oxide for Use in Aqueous Solution”, *2D Materials* **2020**, *7*, 025025.
- [123] T. Murtola, E. Falck, M. Karttunen, I. Vattulainen, “Coarse-Grained Model for Phospholipid/Cholesterol Bilayer Employing Inverse Monte Carlo with Thermodynamic Constraints”, *The Journal of Chemical Physics* **2007**, *126*, 075101.
- [124] S. Mortezaazadeh, Y. Jamali, H. Naderi-Manesh, A. P. Lyubartsev, “Implicit Solvent Systematic Coarse-Graining of Dioleoylphosphatidylethanolamine Lipids: From the Inverted Hexagonal to the Bilayer Structure”, *PLOS ONE* **2019**, *14*, e0214673.
- [125] A. J. Clark, J. McCarty, M. G. Guenza, “Effective Potentials for Representing Polymers in Melts as Chains of Interacting Soft Particles”, *The Journal of Chemical Physics* **2013**, *139*, 124906.
- [126] M. G. Guenza, “Structural and Thermodynamic Consistency in Coarse-Grained Models of Macromolecules”, *Journal of Physics: Conference Series* **2015**, *640*, 012009.
- [127] M. Heinen, “Calculating Particle Pair Potentials from Fluid-State Pair Correlations: Iterative Ornstein–Zernike Inversion”, *Journal of Computational Chemistry* **2018**, *39*, 1531–1543.
- [128] J. Ghosh, R. Faller, “State Point Dependence of Systematically Coarse-Grained Potentials”, *Molecular Simulation* **2007**, *33*, 759–767.
- [129] T. C. Moore, C. R. Iacovella, C. McCabe, “Derivation of Coarse-Grained Potentials via Multistate Iterative Boltzmann Inversion”, *The Journal of Chemical Physics* **2014**, *140*, 224104.
- [130] A. Mirzoev, A. P. Lyubartsev, “MagiC: Software Package for Multiscale Modeling”, *Journal of Chemical Theory and Computation* **2013**, *9*, 1512–1520.
- [131] M. R. DeLyser, W. G. Noid, “Extending Pressure-Matching to Inhomogeneous Systems via Local-Density Potentials”, *The Journal of Chemical Physics* **2017**, *147*, 134111.
- [132] A. Moradzadeh, M. H. Motevaselian, S. Y. Mashayak, N. R. Aluru, “Coarse-Grained Force Field for Imidazolium-Based Ionic Liquids”, *Journal of Chemical Theory and Computation* **2018**, *14*, 3252–3261.
- [133] P. Ganguly, D. Mukherji, C. Junghans, N. F. A. van der Vegt, “Kirkwood–Buff Coarse-Grained Force Fields for Aqueous Solutions”, *Journal of Chemical Theory and Computation* **2012**, *8*, 1802–1807.
- [134] T. E. de Oliveira, P. A. Netz, K. Kremer, C. Junghans, D. Mukherji, “C-IBI: Targeting Cumulative Coordination within an Iterative Protocol to Derive Coarse-Grained Models of (Multi-Component) Complex Fluids”, *The Journal of Chemical Physics* **2016**, *144*, 174106.
- [135] K. S. Schweizer, J. G. Curro in *Atomistic Modeling of Physical Properties*, (Eds.: L. Monnerie, U. W. Suter), Advances in Polymer Science, Springer, Berlin, Heidelberg, **1994**, pp. 319–377.
- [136] S. Y. Mashayak, M. N. Jochum, K. Koschke, N. R. Aluru, V. Rühle, C. Junghans, “Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA”, *PLOS ONE* **2015**, *10*, e0131754.

- 
- 
- [137] W. L. Jorgensen, J. Tirado-Rives, “The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin”, *Journal of the American Chemical Society* **1988**, *110*, 1657–1666.
- [138] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, W. L. Jorgensen, “LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands”, *Nucleic Acids Research* **2017**, *45*, W331–W336.
- [139] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, “GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers”, *SoftwareX* **2015**, *1–2*, 19–25.
- [140] M. Kohns, S. Reiser, M. Horsch, H. Hasse, “Solvent Activity in Electrolyte Solutions from Molecular Simulation of the Osmotic Pressure”, *The Journal of Chemical Physics* **2016**, *144*, 084112.
- [141] P. T. Cummings, G. Stell, “Interaction Site Models for Molecular Fluids”, *Molecular Physics* **1982**, *46*, 383–426.
- [142] M. Kohns, M. Horsch, H. Hasse, “Activity Coefficients from Molecular Simulations Using the OPAS Method”, *The Journal of Chemical Physics* **2017**, *147*, 144108.
- [143] M. Jochum, D. Andrienko, K. Kremer, C. Peter, “Structure-Based Coarse-Graining in Liquid Slabs”, *The Journal of Chemical Physics* **2012**, *137*, 064102.
- [144] S. Dhamankar, M. A. Webb, “Chemically Specific Coarse-Graining of Polymers: Methods and Prospects”, *Journal of Polymer Science* **2021**, *59*, 2613–2643.
- [145] A. E. Hafner, J. Krausser, A. Šarić, “Minimal Coarse-Grained Models for Molecular Self-Organisation in Biology”, *Current Opinion in Structural Biology*, Cryo Electron Microscopy - Biophysical and Computational Methods - Biophysical and Computational Methods - Part B **2019**, *58*, 43–52.
- [146] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, C. Clementi, “Machine Learning of Coarse-Grained Molecular Dynamics Force Fields”, *ACS Central Science* **2019**, *5*, 755–767.
- [147] M. Hanke, *A Taste of Inverse Problems*, Society for Industrial and Applied Mathematics, **2017**, 148 pp.
- [148] M. P. Bernhardt, M. Hanke, N. F. A. van der Vegt, “Iterative Integral Equation Methods for Structural Coarse-Graining”, *The Journal of Chemical Physics* **2021**, *154*, 084118.
- [149] H. Bertagnolli, I. Hausleithner, O. Steinhauser, “Symmetry Reduction of the RISM Equation”, *Chemical Physics Letters* **1985**, *116*, 465–470.
- [150] E. Pretti, M. S. Shell, “A Microcanonical Approach to Temperature-Transferable Coarse-Grained Models Using the Relative Entropy”, *The Journal of Chemical Physics* **2021**, *155*, 094102.
- [151] C.-C. Fu, P. M. Kulkarni, M. Scott Shell, L. Gary Leal, “A Test of Systematic Coarse-Graining of Molecular Dynamics Simulations: Thermodynamic Properties”, *The Journal of Chemical Physics* **2012**, *137*, 164106.

- 
- [152] M. P. Bernhardt, Y. Nagata, N. F. A. van der Vegt, "Where Lennard-Jones Potentials Fail: Iterative Optimization of Ion–Water Pair Potentials Based on Ab Initio Molecular Dynamics Data", *The Journal of Physical Chemistry Letters* **2022**, 3712–3717.
- [153] G. Bussi, D. Donadio, M. Parrinello, "Canonical Sampling through Velocity Rescaling", *The Journal of Chemical Physics* **2007**, *126*, 014101.
- [154] L. Verlet, J.-J. Weis, "Perturbation Theory for the Thermodynamic Properties of Simple Liquids", *Molecular Physics* **1972**, *24*, 1013–1024.
- [155] J. O. Hirschfelder, C. F. Curtis, R. Bird, *Molecular Theory of Gases and Liquids*, corrected, John Wiley and Sons, New York, **1964**.
- [156] W. L. Jorgensen, J. Tirado-Rives, "Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems", *Proceedings of the National Academy of Sciences* **2005**, *102*, 6665–6670.
- [157] L. S. Dodda, J. Z. Vilseck, J. Tirado-Rives, W. L. Jorgensen, "1.14\*CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations", *The Journal of Physical Chemistry B* **2017**, *121*, 3864–3870.
- [158] A. Lyubartsev, Y. Tu, A. Laaksonen, "Hierarchical Multiscale Modelling Scheme from First Principles to Mesoscale", *Journal of Computational and Theoretical Nanoscience* **2009**, *6*, 951–959.
- [159] A. Mirzoev, L. Nordenskiöld, A. Lyubartsev, "Magic v.3: An Integrated Software Package for Systematic Structure-Based Coarse-Graining", *Computer Physics Communications* **2019**, *237*, 263–273.
- [160] W. R. Smith, I. Nezbeda, J. Kolafa, F. Moučka, "Recent Progress in the Molecular Simulation of Thermodynamic Properties of Aqueous Electrolyte Solutions", *Fluid Phase Equilibria* **2018**, *466*, 19–30.
- [161] I. Nezbeda, F. Moučka, W. R. Smith, "Recent Progress in Molecular Simulation of Aqueous Electrolytes: Force Fields, Chemical Potentials and Solubility", *Molecular Physics* **2016**, *114*, 1665–1690.
- [162] M. Fyta, I. Kalcher, J. Dzubiella, L. Vrbka, R. R. Netz, "Ionic Force Field Optimization Based on Single-Ion and Ion-Pair Solvation Properties", *The Journal of Chemical Physics* **2010**, *132*, 024911.
- [163] E. Wernersson, P. Jungwirth, "Effect of Water Polarizability on the Properties of Solutions of Polyvalent Ions: Simulations of Aqueous Sodium Sulfate with Different Force Fields", *Journal of Chemical Theory and Computation* **2010**, *6*, 3233–3240.
- [164] D. A. Tolmachev, O. S. Boyko, N. V. Lukasheva, H. Martinez-Seara, M. Karttunen, "Overbinding and Qualitative and Quantitative Changes Caused by Simple Na<sup>+</sup> and K<sup>+</sup> Ions in Polyelectrolyte Simulations: Comparison of Force Fields with and without NBFIX and ECC Corrections", *Journal of Chemical Theory and Computation* **2020**, *16*, 677–687.
- [165] D. E. Smith, L. X. Dang, "Computer Simulations of NaCl Association in Polarizable Water", *The Journal of Chemical Physics* **1994**, *100*, 3757–3766.

- 
- [166] F. Sebastiani, A. V. Verde, M. Heyden, G. Schwaab, M. Havenith, “Cooperativity and Ion Pairing in Magnesium Sulfate Aqueous Solutions from the Dilute Regime to the Solubility Limit”, *Physical Chemistry Chemical Physics* **2020**, *22*, 12140–12153.
- [167] L. Pegado, O. Marsalek, P. Jungwirth, E. Wernersson, “Solvation and Ion-Pairing Properties of the Aqueous Sulfate Anion: Explicit versus Effective Electronic Polarization”, *Physical Chemistry Chemical Physics* **2012**, *14*, 10248–10257.
- [168] R. Fuentes-Azcatl, M. C. Barbosa, “Sodium Chloride, NaCl/ $\epsilon$ : New Force Field”, *The Journal of Physical Chemistry B* **2016**, *120*, 2460–2470.
- [169] M. Předota, D. Biriukov, “Electronic Continuum Correction without Scaled Charges”, *Journal of Molecular Liquids* **2020**, *314*, 113571.
- [170] S. Weerasinghe, P. E. Smith, “A Kirkwood–Buff Derived Force Field for Sodium Chloride in Water”, *The Journal of Chemical Physics* **2003**, *119*, 11342–11349.
- [171] B. Hess, N. F. A. van der Vegt, “Cation Specific Binding with Protein Surface Charges”, *Proceedings of the National Academy of Sciences* **2009**, *106*, 13296–13300.
- [172] S. Mamatkulov, M. Fyta, R. R. Netz, “Force Fields for Divalent Cations Based on Single-Ion and Ion-Pair Properties”, *The Journal of Chemical Physics* **2013**, *138*, 024505.
- [173] J. L. Aragones, M. Rovere, C. Vega, P. Gallo, “Computer Simulation Study of the Structure of LiCl Aqueous Solutions: Test of Non-Standard Mixing Rules in the Ion Interaction”, *The Journal of Physical Chemistry B* **2014**, *118*, 7680–7691.
- [174] P. Li, B. P. Roberts, D. K. Chakravorty, K. M. Merz, “Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent”, *Journal of Chemical Theory and Computation* **2013**, *9*, 2733–2748.
- [175] A. Chaumont, E. Engler, R. Schurhammer, “Is Charge Scaling Really Mandatory When Developing Fixed-Charge Atomistic Force Fields for Deep Eutectic Solvents?”, *The Journal of Physical Chemistry B* **2020**, *124*, 7239–7250.
- [176] P. Loche, P. Steinbrunner, S. Friedowitz, R. R. Netz, D. J. Bonthuis, “Transferable Ion Force Fields in Water from a Simultaneous Optimization of Ion Solvation and Ion–Ion Interaction”, *The Journal of Physical Chemistry B* **2021**, *125*, 8581–8587.
- [177] X. Periole, D. Allouche, J.-P. Daudey, Y.-H. Sanejouand, “Simple Two-Body Cation–Water Interaction Potentials Derived from *Ab Initio* Calculations. Comparison to Results Obtained with an Empirical Approach”, *The Journal of Physical Chemistry B* **1997**, *101*, 5018–5025.
- [178] X. Periole, D. Allouche, A. Ramírez-Solís, I. Ortega-Blake, J. P. Daudey, Y. H. Sanejouand, “New Effective Potentials Extraction Method for the Interaction between Cations and Water”, *The Journal of Physical Chemistry B* **1998**, *102*, 8579–8587.
- [179] K. Kasahara, Y. Takimoto, R. Ashida, T. Takahashi, “Effects of Ion–Water Lennard-Jones Potentials on the Hydration Dynamics around a Monovalent Atomic Ion in Molecular Dynamics Simulations”, *Molecular Simulation* **2019**, *46*, 83–91.
- [180] S. Gruenhut, D. R. MacFarlane, “Molecular Dynamics Simulation of Heavy Metal Fluoride Glasses: Comparison of Buckingham and BHM Potentials”, *Journal of Non-Crystalline Solids, Non-Oxide Glasses* **1995**, *184*, 356–362.

- 
- [181] V. R. Manga, S. Bringuier, J. Paul, S. Jayaraman, P. Lucas, P. Deymier, K. Muralidharan, "Molecular Dynamics Simulations and Thermodynamic Modeling of NaCl–KCl–ZnCl<sub>2</sub> Ternary System", *CALPHAD* **2014**, *46*, 176–183.
- [182] V. Migliorati, A. Serva, F. M. Terenzio, P. D'Angelo, "Development of Lennard-Jones and Buckingham Potentials for Lanthanoid Ions in Water", *Inorganic Chemistry* **2017**, *56*, 6214–6224.
- [183] M.-M. Walz, M. M. Ghahremanpour, P. J. van Maaren, D. van der Spoel, "Phase-Transferable Force Field for Alkali Halides", *Journal of Chemical Theory and Computation* **2018**, *14*, 5933–5948.
- [184] D. A. Schmidt, Ö. Birer, S. Funkner, B. P. Born, R. Gnanasekaran, G. W. Schwaab, D. M. Leitner, M. Havenith, "Rattling in the Cage: Ions as Probes of Sub-picosecond Water Network Dynamics", *Journal of the American Chemical Society* **2009**, *131*, 18512–18517.
- [185] S. Funkner, G. Niehues, D. A. Schmidt, M. Heyden, G. Schwaab, K. M. Callahan, D. J. Tobias, M. Havenith, "Watching the Low-Frequency Motions in Aqueous Salt Solutions: The Terahertz Vibrational Signatures of Hydrated Ions", *Journal of the American Chemical Society* **2012**, *134*, 1030–1035.
- [186] G. Schwaab, F. Sebastiani, M. Havenith, "Ion Hydration and Ion Pairing as Probed by THz Spectroscopy", *Angewandte Chemie International Edition* **2019**, *58*, 3000–3013.
- [187] W. Schommers, "Pair Potentials in Disordered Many-Particle Systems: A Study for Liquid Gallium", *Physical Review A* **1983**, *28*, 3599–3605.
- [188] J. Chandrasekhar, D. C. Spellmeyer, W. L. Jorgensen, "Energy Component Analysis for Dilute Aqueous Solutions of Lithium(1+), Sodium(1+), Fluoride(1-), and Chloride(1-) Ions", *Journal of the American Chemical Society* **1984**, *106*, 903–910.
- [189] J. Åqvist, "Ion-Water Interaction Potentials Derived from Free Energy Perturbation Simulations", *The Journal of Physical Chemistry* **1990**, *94*, 8021–8024.
- [190] M. Kohagen, P. E. Mason, P. Jungwirth, "Accounting for Electronic Polarization Effects in Aqueous Sodium Chloride via Molecular Dynamics Aided by Neutron Scattering", *The Journal of Physical Chemistry B* **2016**, *120*, 1454–1460.
- [191] T. Martinek, E. Duboué-Dijon, Š. Timr, P. E. Mason, K. Baxová, H. E. Fischer, B. Schmidt, E. Pluhařová, P. Jungwirth, "Calcium Ions in Aqueous Solutions: Accurate Force Field Description Aided by Ab Initio Molecular Dynamics and Neutron Scattering", *The Journal of Chemical Physics* **2018**, *148*, 222813.
- [192] J. L. F. Abascal, C. Vega, "A General Purpose Model for the Condensed Phases of Water: TIP4P/2005", *The Journal of Chemical Physics* **2005**, *123*, 234505.
- [193] A. A. Khan, "Radial Distribution Functions of Fluid Argon", *Physical Review* **1964**, *134*, A367–A384.
- [194] T. Yagasaki, M. Matsumoto, H. Tanaka, "Lennard-Jones Parameters Determined to Reproduce the Solubility of NaCl and KCl in SPC/E, TIP3P, and TIP4P/2005 Water", *Journal of Chemical Theory and Computation* **2020**, *16*, 2460–2473.

- 
- 
- [195] S.-T. Lin, M. Blanco, W. A. G. Iii, "The Two-Phase Model for Calculating Thermodynamic Properties of Liquids from Molecular Dynamics: Validation for the Phase Diagram of Lennard-Jones Fluids", *The Journal of Chemical Physics* **2003**, *119*, 11792–11805.
- [196] S.-T. Lin, P. K. Maiti, W. A. Goddard, "Two-Phase Thermodynamic Model for Efficient and Accurate Absolute Entropy of Water from Molecular Dynamics Simulations", *The Journal of Physical Chemistry B* **2010**, *114*, 8191–8198.
- [197] J. S. Kim, Z. Wu, A. R. Morrow, A. Yethiraj, A. Yethiraj, "Self-Diffusion and Viscosity in Electrolyte Solutions", *The Journal of Physical Chemistry B* **2012**, *116*, 12007–12013.
- [198] Y. Ding, A. A. Hassanali, M. Parrinello, "Anomalous Water Diffusion in Salt Solutions", *Proceedings of the National Academy of Sciences* **2014**, *111*, 3310–3315.
- [199] Z. R. Kann, J. L. Skinner, "A Scaled-Ionic-Charge Simulation Model That Reproduces Enhanced and Suppressed Water Diffusion in Aqueous Salt Solutions", *The Journal of Chemical Physics* **2014**, *141*, 104507.
- [200] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khatullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. TAILLEFUMIER, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, J. Hutter, "CP2K: An Electronic Structure and Molecular Dynamics Software Package - Quickstep: Efficient and Accurate Electronic Structure Calculations", *The Journal of Chemical Physics* **2020**, *152*, 194103.
- [201] Y. Zhang, W. Yang, "Comment on "Generalized Gradient Approximation Made Simple"", *Physical Review Letters* **1998**, *80*, 890–890.
- [202] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, "A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu", *The Journal of Chemical Physics* **2010**, *132*, 154104.
- [203] S. Goedecker, M. Teter, J. Hutter, "Separable Dual-Space Gaussian Pseudopotentials", *Physical Review B* **1996**, *54*, 1703–1710.
- [204] L. Verlet, "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules", *Physical Review* **1967**, *159*, 98–103.
- [205] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, W. F. van Gunsteren, "Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations", *Chemical Physics Letters* **1994**, *222*, 529–539.
- [206] G. Hummer, L. R. Pratt, A. E. García, B. J. Berne, S. W. Rick, "Electrostatic Potentials and Free Energies of Solvation of Polar and Charged Molecules", *The Journal of Physical Chemistry B* **1997**, *101*, 3017–3020.
- [207] I. V. Leontyev, A. A. Stuchebrukhov, "Dielectric Relaxation of Cytochrome c Oxidase: Comparison of the Microscopic and Continuum Models", *The Journal of Chemical Physics* **2009**, *130*, 085103.
- [208] Y. Marcus, "Ionic Radii in Aqueous Solutions", *Chemical Reviews* **1988**, *88*, 1475–1498.

- 
- [209] D. M. de Oliveira, S. R. Zukowski, V. Palivec, J. Hénin, H. Martinez-Seara, D. Ben-Amotz, P. Jungwirth, E. Duboué-Dijon, “Binding of Divalent Cations to Acetate: Molecular Simulations Guided by Raman Spectroscopy”, *Physical Chemistry Chemical Physics* **2020**, *22*, 24014–24027.
- [210] B. Hess, N. F. A. van der Vegt, “Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models”, *The Journal of Physical Chemistry B* **2006**, *110*, 17616–17626.
- [211] J. Ángel Sordo, “Solvation Thermodynamics: Two Formulations and Some Misunderstandings”, *RSC Advances* **2015**, *5*, 96105–96116.
- [212] D. W. Green, R. H. Perry, *Perry’s Chemical Engineers’ Handbook, Eighth Edition*, McGraw-Hill Education, **2008**.
- [213] B. Dünweg, K. Kremer, “Molecular Dynamics Simulation of a Polymer Chain in Solution”, *The Journal of Chemical Physics* **1993**, *99*, 6983–6997.
- [214] M. A. González, J. L. F. Abascal, “The Shear Viscosity of Rigid Water Models”, *The Journal of Chemical Physics* **2010**, *132*, 096101.
- [215] K. J. Müller, H. G. Hertz, “A Parameter as an Indicator for Water–Water Association in Solutions of Strong Electrolytes”, *The Journal of Physical Chemistry* **1996**, *100*, 1256–1265.
- [216] M. E. Guendouzi, A. Dinane, A. Mounir, “Water Activities, Osmotic and Activity Coefficients in Aqueous Chloride Solutions at  $T = 298.15$  K by the Hygrometric Method”, *The Journal of Chemical Thermodynamics* **2001**, *33*, 1059–1072.
- [217] W. M. Haynes, *CRC Handbook of Chemistry and Physics*, CRC Press, **2016**, 2643 pp.
- [218] M. G. Saunders, G. A. Voth, “Coarse-Graining Methods for Computational Biology”, *Annual Review of Biophysics* **2013**, *42*, 73–93.
- [219] D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, K. Kremer, “Multiscale Modeling of Soft Matter: Scaling of Dynamics”, *Physical Chemistry Chemical Physics* **2011**, *13*, 10412–10420.
- [220] R. M. Fuchsli, H. Fellermann, A. Eriksson, H.-J. Ziock, “Coarse Graining and Scaling in Dissipative Particle Dynamics”, *The Journal of Chemical Physics* **2009**, *130*, 214102.
- [221] S. Izvekov, B. M. Rice, “Multi-Scale Coarse-Graining of Non-Conservative Interactions in Molecular Liquids”, *The Journal of Chemical Physics* **2014**, *140*, 104104.
- [222] S. Trément, B. Schnell, L. Petitjean, M. Couty, B. Rousseau, “Conservative and Dissipative Force Field for Simulation of Coarse-Grained Alkane Molecules: A Bottom-up Approach”, *The Journal of Chemical Physics* **2014**, *140*, 134113.
- [223] G. Deichmann, N. F. A. van der Vegt, “Bottom-up Approach to Represent Dynamic Properties in Coarse-Grained Molecular Simulations”, *The Journal of Chemical Physics* **2018**, *149*, 244114.
- [224] J. F. Rudzinski, “Recent Progress towards Chemically-Specific Coarse-Grained Simulation Models with Consistent Dynamical Properties”, *Computation* **2019**, *7*, 42.

- 
- 
- [225] Y. Rosenfeld, "Relation between the Transport Coefficients and the Internal Entropy of Simple Systems", *Physical Review A* **1977**, *15*, 2545–2549.
- [226] J. Mittal, J. R. Errington, T. M. Truskett, "Relationships between Self-Diffusivity, Packing Fraction, and Excess Entropy in Simple Bulk and Confined Fluids", *The Journal of Physical Chemistry B* **2007**, *111*, 10054–10063.
- [227] B. Mukherjee, C. Peter, K. Kremer, "Single Molecule Translocation in Smectics Illustrates the Challenge for Time-Mapping in Simulations on Multiple Scales", *The Journal of Chemical Physics* **2017**, *147*, 114501.
- [228] J. Jin, Y. Han, G. A. Voth, "Coarse-Graining Involving Virtual Sites: Centers of Symmetry Coarse-Graining", *The Journal of Chemical Physics* **2019**, *150*, 154103.
- [229] D. Fritz, V. A. Harmandaris, K. Kremer, N. F. A. van der Vegt, "Coarse-Grained Polymer Melts Based on Isolated Atomistic Chains: Simulation of Polystyrene of Different Tacticities", *Macromolecules* **2009**, *42*, 7579–7588.
- [230] J. F. Rudzinski, T. Bereau, "Structural-Kinetic-Thermodynamic Relationships Identified from Physics-Based Molecular Simulation Models", *The Journal of Chemical Physics* **2018**, *148*, 204111.
- [231] R. Baron, A. H. de Vries, P. H. Hünenberger, W. F. van Gunsteren, "Comparison of Atomic-Level and Coarse-Grained Models for Liquid Hydrocarbons from Molecular Dynamics Configurational Entropy Estimates", *The Journal of Physical Chemistry B* **2006**, *110*, 8464–8473.
- [232] A. Chaimovich, M. S. Shell, "Anomalous Waterlike Behavior in Spherically-Symmetric Water Models Optimized with the Relative Entropy", *Physical Chemistry Chemical Physics* **2009**, *11*, 1901–1915.
- [233] S. Riniker, J. R. Allison, W. F. van Gunsteren, "On Developing Coarse-Grained Models for Biomolecular Simulation: A Review", *Physical chemistry chemical physics: PCCP* **2012**, *14*, 12423–12430.
- [234] S. Kmiciek, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, A. Kolinski, "Coarse-Grained Protein Models and Their Applications", *Chemical Reviews* **2016**, *116*, 7898–7936.
- [235] R. Baron, A. H. de Vries, P. H. Hünenberger, W. F. van Gunsteren, "Configurational Entropies of Lipids in Pure and Mixed Bilayers from Atomic-Level and Coarse-Grained Molecular Dynamics Simulations", *The Journal of Physical Chemistry B* **2006**, *110*, 15602–15614.
- [236] J. F. Rudzinski, W. G. Noid, "Coarse-Graining Entropy, Forces, and Structures", *The Journal of Chemical Physics* **2011**, *135*, 214101.
- [237] J. Jin, A. J. Pak, G. A. Voth, "Understanding Missing Entropy in Coarse-Grained Systems: Addressing Issues of Representability and Transferability", *The Journal of Physical Chemistry Letters* **2019**, *10*, 4549–4557.
- [238] G. Deichmann, N. F. A. van der Vegt, "Conditional Reversible Work Coarse-Grained Models of Molecular Liquids with Coulomb Electrostatics – A Proof of Concept Study on Weakly Polar Organic Molecules", *Journal of Chemical Theory and Computation* **2017**, *13*, 6158–6166.

- 
- 
- [239] T. A. Pascal, W. A. Goddard, Y. Jung, "Entropy and the Driving Force for the Filling of Carbon Nanotubes with Water", *Proceedings of the National Academy of Sciences* **2011**, *108*, 11794–11798.
- [240] T. A. Pascal, S.-T. Lin, W. A. G. Iii, "Thermodynamics of Liquids: Standard Molar Entropies and Heat Capacities of Common Solvents from 2PT Molecular Dynamics", *Physical Chemistry Chemical Physics* **2010**, *13*, 169–181.
- [241] N. F. Carnahan, K. E. Starling, "Equation of State for Nonattracting Rigid Spheres", *The Journal of Chemical Physics* **1969**, *51*, 635–636.
- [242] P.-K. Lai, S.-T. Lin, "Rapid Determination of Entropy for Flexible Molecules in Condensed Phase from the Two-Phase Thermodynamic Model", *RSC Advances* **2014**, *4*, 9522–9533.
- [243] S.-N. Huang, T. A. Pascal, W. A. Goddard, P. K. Maiti, S.-T. Lin, "Absolute Entropy and Energy of Carbon Dioxide Using the Two-Phase Thermodynamic Model", *Journal of Chemical Theory and Computation* **2011**, *7*, 1893–1901.
- [244] M. A. Caro, T. Laurila, O. Lopez-Acevedo, "Accurate Schemes for Calculation of Thermodynamic Properties of Liquid Mixtures from Molecular Dynamics Simulations", *The Journal of Chemical Physics* **2016**, *145*, 244504.
- [245] M. Bernhardt, MA thesis, TU Darmstadt, **2016**.
- [246] C. Eckart, "Some Studies Concerning Rotating Axes and Polyatomic Molecules", *Physical Review* **1935**, *47*, 552–558.
- [247] J. D. Louck, H. W. Galbraith, "Eckart Vectors, Eckart Frames, and Polyatomic Molecules", *Reviews of Modern Physics* **1976**, *48*, 69–106.
- [248] J. Sablić, R. Delgado-Buscalioni, M. Praprotnik, "Application of the Eckart Frame to Soft Matter: Rotation of Star Polymers under Shear Flow", *Soft Matter* **2017**, *13*, 6988–7000.
- [249] Y. M. Rhee, M. S. Kim, "Mode-Specific Energy Analysis for Rotating-Vibrating Triatomic Molecules in Classical Trajectory Simulation", *The Journal of Chemical Physics* **1997**, *107*, 1394–1402.
- [250] C. Heidelbach, V. S. Vikhrenko, D. Schwarzer, J. Schroeder, "Molecular Dynamics Simulation of Vibrational Relaxation of Highly Excited Molecules in Fluids. II. Nonequilibrium Simulation of Azulene in CO<sub>2</sub> and Xe", *The Journal of Chemical Physics* **1999**, *110*, 5286–5299.
- [251] M. L. P. Price, D. Ostrovsky, W. L. Jorgensen, "Gas-Phase and Liquid-State Properties of Esters, Nitriles, and Nitro Compounds with the OPLS-AA Force Field", *Journal of Computational Chemistry* **2001**, *22*, 1340–1352.
- [252] T. Darden, D. York, L. Pedersen, "Particle Mesh Ewald: An N log(N) Method for Ewald Sums in Large Systems", *Journal of Chemical Physics* **1993**, *98*, 10089.
- [253] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, "A Smooth Particle Mesh Ewald Method", *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.
- [254] S. Nosé, "A Unified Formulation of the Constant Temperature Molecular Dynamics Methods", *The Journal of Chemical Physics* **1984**, *81*, 511–519.

- 
- 
- [255] M. Parrinello, A. Rahman, “Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method”, *Journal of Applied Physics* **1981**, 52, 7182–7190.
- [256] G. Deichmann, V. Marcon, N. F. A. van der Vegt, “Bottom-up Derivation of Conservative and Dissipative Interactions for Coarse-Grained Molecular Liquids with the Conditional Reversible Work Method”, *The Journal of Chemical Physics* **2014**, 141, 224109.
- [257] E. Brini, N. F. A. van der Vegt, “Chemically Transferable Coarse-Grained Potentials from Conditional Reversible Work Calculations”, *The Journal of Chemical Physics* **2012**, 137, 154113.
- [258] V. R. Ardham, G. Deichmann, N. F. A. van der Vegt, F. Leroy, “Solid-Liquid Work of Adhesion of Coarse-Grained Models of n-Hexane on Graphene Layers Derived from the Conditional Reversible Work Method”, *The Journal of Chemical Physics* **2015**, 143, 243135.
- [259] S. Plimpton, “Fast Parallel Algorithms for Short-Range Molecular Dynamics”, *Journal of Computational Physics* **1995**, 117, 1–19.
- [260] G. J. Martyna, D. J. Tobias, M. L. Klein, “Constant Pressure Molecular Dynamics Algorithms”, *The Journal of Chemical Physics* **1994**, 101, 4177–4189.
- [261] M. E. Tuckerman, J. Alejandre, R. López-Rendón, A. L. Jochim, G. J. Martyna, “A Liouville-operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in the Isothermal–Isobaric Ensemble”, *Journal of Physics A: Mathematical and General* **2006**, 39, 5629–5651.
- [262] R. W. Hockney, J. W. Eastwood, “Computer Simulation Using Particles”, *Bristol: Hilger* **1988**.
- [263] H. A. Karimi-Varzaneh, N. F. A. van der Vegt, F. Müller-Plathe, P. Carbone, “How Good Are Coarse-Grained Polymer Models? A Comparison for Atactic Polystyrene”, *ChemPhysChem* **2012**, 13, 3428–3439.
- [264] Z. Li, X. Bian, X. Li, G. E. Karniadakis, “Incorporation of Memory Effects in Coarse-Grained Modeling via the Mori-Zwanzig Formalism”, *The Journal of Chemical Physics* **2015**, 143, 243128.
- [265] J. C. Dyre, “Perspective: Excess-entropy Scaling”, *The Journal of Chemical Physics* **2018**, 149, 210901.
- [266] J. A. Armstrong, C. Chakravarty, P. Ballone, “Statistical Mechanics of Coarse Graining: Estimating Dynamical Speedups from Excess Entropies”, *The Journal of Chemical Physics* **2012**, 136, 124503.
- [267] M. S. Shell, “Systematic Coarse-Graining of Potential Energy Landscapes and Dynamics in Liquids”, *The Journal of Chemical Physics* **2012**, 137, 084503.
- [268] O. Sackur, “Die Anwendung Der Kinetischen Theorie Der Gase Auf Chemische Probleme”, *Annalen der Physik* **1911**, 341, 958–980.
- [269] R. E. Nettleton, M. S. Green, “Expression in Terms of Molecular Distribution Functions for the Entropy Density in an Infinite System”, *The Journal of Chemical Physics* **1958**, 29, 1365–1370.

- 
- 
- [270] E. Voyiatzis, F. Müller-Plathe, M. C. Böhm, “Do Transport Properties of Entangled Linear Polymers Scale with Excess Entropy?”, *Macromolecules* **2013**, *46*, 8710–8723.
- [271] N. Jakse, A. Pasturel, “Excess Entropy Scaling Law for Diffusivity in Liquid Metals”, *Scientific Reports* **2016**, *6*, 1–11.
- [272] R. Chopra, T. M. Truskett, J. R. Errington, “On the Use of Excess Entropy Scaling To Describe Single-Molecule and Collective Dynamic Properties of Hydrocarbon Isomer Fluids”, *The Journal of Physical Chemistry B* **2010**, *114*, 16487–16493.
- [273] E. W. Lemmon, I. H. Bell, M. L. Huber, M. O. McLinden in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology, Gaithersburg MD, p. 20899.
- [274] K. Shen, N. Sherck, M. Nguyen, B. Yoo, S. Köhler, J. Speros, K. T. Delaney, G. H. Fredrickson, M. S. Shell, “Learning Composition-Transferable Coarse-Grained Models: Designing External Potential Ensembles to Maximize Thermodynamic Information”, *The Journal of Chemical Physics* **2020**, *153*, 154116.
- [275] P. F. McMillan, “Chemistry at High Pressure”, *Chemical Society Reviews* **2006**, *35*, 855–857.
- [276] L. Huang, B. Roux, “Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data”, *Journal of Chemical Theory and Computation* **2013**, *9*, 3543–3556.
- [277] J. Timko, D. Bucher, S. Kuyucak, “Dissociation of NaCl in Water from Ab Initio Molecular Dynamics Simulations”, *The Journal of Chemical Physics* **2010**, *132*, 114510.
- [278] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, “Machine Learning of Accurate Energy-Conserving Molecular Force Fields”, *Science Advances* **2017**, *3*, e1603015.
- [279] J. Zavadlav, S. Bevc, M. Praprotnik, “Adaptive Resolution Simulations of Biomolecular Systems”, *European Biophysics Journal* **2017**, *46*, 821–835.
- [280] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, “Improved Protein Structure Prediction Using Potentials from Deep Learning”, *Nature* **2020**, *577*, 706–710.
- [281] I. Lyubimov, M. G. Guenza, “First-Principle Approach to Rescale the Dynamics of Simulated Coarse-Grained Macromolecular Liquids”, *Physical Review E* **2011**, *84*, 031801.