Disentangling solvation and structural response in quantitative calculations of spectroscopic parameters

Die Dissertation wurde im Zeitraum 11.2020 bis 04.2026 angefertigt und der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund vorgelegt

von Stefan Maste

Dortmund 2026

Erstgutachter: Prof. Stefan M. Kast

Zweitgutachter: Prof. Roland Winter

Danksagung

Zuerst einmal danke ich Prof. Dr. Stefan Kast für die Betreuung dieser Promotion, für die Möglichkeit mich im Rahmen meiner Forschung, sowohl während der Promotion als auch vorherigen Abschlussarbeiten frei entfalten zu können und somit meine bisherige wissenschaftliche Karriere ermöglicht zu haben.

Ich danke Prof. Dr. Roland Winter für die Übernahme des Zweitgutachtens und für den spannenden wissenschaftlichen und nicht-wissenschaftlichen Austausch im Rahmen von RESOLV auf Klausurtagungen.

Dem gesamten Arbeitskreis danke ich für die schöne Arbeitsatmosphäre und ständige Hilfsbereitschaft, die auch in Zeiten von Corona und Home-Office geherrscht hat. Hinzukommen die zahlreichen Tagungen, die wir zusammen besucht haben und auf denen wir viele lustige Momente zusammen erleben konnten.

Zusätzlich danke ich den vielen Menschen mit denen ich an den diversen Projekten arbeiten konnte. Ich danke insbesondere Prof. Dr. Wolf Hiller, Dr. Bastian Grabe und Benjamin Kissel für die Zusammenarbeit bei den Hochdruck-NMR-Messungen. I thank Prof. Dr. Enrica Bordignon and Dr. Laura Galazzo for the collaboration in the EPR project. I thank Prof. Dr. Dominik Marx and Dr. Bikramjit Sharma for the collaboration in the NMR and EPR project. Außerdem danke ich Dr. Tim Pongratz für die Zusammenarbeit im NMR und EPR Projekt und Michael Strobl für die Zusammenarbeit im Machine Learning Projekt.

I thank the whole group of Prof. Adrian Roitberg for the welcoming atmosphere during my research internship and especially Kate Huddleston, Ignacio Pickering and Jinze Xue for the collaboration in the Machine Learning project.

Ich danke dem Excellenzcluster RESOLV, die mir den Besuch der zahlreichen Tagungen und den Forschungsaufenthalt in den USA ermöglicht haben.

Zuletzt danke ich meinen Eltern, die mich bereits mein gesamtes Leben unterstützt haben und mir immer alles ermöglicht haben und Laura, die mich während der gesamten Promotion durch alle Höhen und Tiefen begleitet hat und immer an meiner Seite war. Ohne euch hätte ich das alles nicht geschafft.

Publications

Parts of this thesis were already published in the following publications:

S. Maste, B. Sharma, T. Pongratz, B. Grabe, W. Hiller, M. B. Erlach, W. Kremer, H. R. Kalbitzer, D. Marx, S. M. Kast, *Phys. Chem. Chem. Phys.* **2024**, 6386. DOI: 10.1039/D3CP05471C

Electronic research data: DOI: 10.22000/1743

L. Galazzo, S. Maste, B. Sharma, A. van Tran, T. Pongratz, M. Teucher, D. Marx, F. Neese, S. M. Kast, E. Bordignon, *J. Phys. Chem. Lett.* **2025**, 8141. DOI: 10.1021/acs.jpclett.5c01053

Electronic research data: DOI: 10.17877/RESOLV-2025-M7UFWH5H

Table of Contents

Zusammenfassung

Das Verständnis von Prozessen auf molekularer und sogar atomarer Ebene ist das zentrale Element der Chemie. Neben experimentellen Techniken wie der Spektroskopie wird dieses Verständnis immer mehr durch den Einsatz von theoretischen Methoden gewonnen. Historisch waren hier vor allem quantenchemischen Rechnungen oder Molekulardynamik(MD)-Simulationen die Methoden der Wahl, um chemische Prozesse tiefer zu verstehen. Insbesondere in den letzten Jahren hat der technische Fortschritt im Umgang mit großen Datenmengen auch Methoden der künstlichen Intelligenz hervorgerufen, die auch in der Chemie Einsatz finden.

Im Rahmen dieser Arbeit wurden all diese Methoden kombiniert, um ein tieferes Verständnis über die Chemie auf atomarer Ebene zu erlangen. Hierbei lag ein besonderer Fokus auf dem Verständnis von Solvatationsbedingungen und deren Veränderung unter extremen Bedingungen wie hohem hydrostatischen Druck. Auf experimenteller Seite wurde die Veränderung von Parametern der Kernmagnetresonanzspektroskopie (NMR) unter hohem hydrostatischem Druck untersucht. Hierfür wurde eine Reihe von kleinen Molekülen vermessen. Diese Experimente wurden genutzt um theoretische Methoden zu optimieren, um das Verhalten unter hohem hydrostatischem Druck richtig abzubilden. Hierfür wurde eine Kombination von quantenchemischen Rechnungen und MD-Simulationen verwendet. Im Rahmen der quantenchemischen Rechnungen wurde das *Embedded Cluster Reference Interaction Site Model* (EC-RISM) genutzt, was in der Lage ist, Hochdruckbedingungen abzubilden. Dies wurde auch genutzt, um die strukturelle Veränderung unter Hochdruckbedingungen abzubilden. Durch den Vergleich mit zuverlässigen experimentellen Daten konnte diese Methode inklusive der daraus abgeleiteten Strukturänderungen validiert werden. Es zeigte sich, dass der Einfluss von hohem Druck auf die Struktur von kleinen Molekülen und auch auf das konformationelle Gleichgewicht von zwei Dipeptiden so gering ist, dass für genaue quantenchemische Vorhersagen von NMR-Parametern diese kleinen Strukturänderungen berücksichtigt werden müssen. Dies konnte auf die hohe strukturelle Sensitivität von NMR-Parametern zurückgeführt werden, wodurch sich die Vorhersage dieser Parameter als ideal für den Vergleich theoretischer Methoden erwiesen hat.

Dies konnte in einem weiteren Projekt für den Vergleich von quantenchemischen *ab initio* (AI)-, klassischen Kraftfeld und *Machine Learning* (ML)-MD-Simulationen genutzt werden. Ausgehend von hochgenauen AIMD-Simulationen wurden verschiedene Methoden getestet das Lösungsmittel im Rahmen von quantenchemischen Rechnungen darzustellen. Durch den Vergleich mit experimentellen NMR-Parametern konnte die Validität der verschiedenen Methoden überprüft werden. Es zeigte sich, dass eine Kombination von expliziten Solvensmolekülen in der Nähe des fokussierten Atoms zusammen mit einer EC-RISM Hintergrundsolvatation sowohl den Rechenaufwand reduziert als auch sehr genaue Vorhersagen liefert. Im Vergleich dazu konnten auf struktureller Ebene des Solvats und in der lokalen Solvatation Abweichungen in der klassischen MD-Simulation der gleichen Moleküle identifiziert werden, die mit einem Fehler in der Vorhersage von NMR-Parametern korrelieren. Im

Vergleich dazu lieferte die ML-MD eine akkurate Darstellung des Solvats, zeigte jedoch Schwächen in der Darstellung des Solvens, die durch eine Erweiterung des Trainingsdatensatzes reduziert werden konnten.

In einem weiteren Projekt dieser Arbeit wurde die pH-Abhängigkeit von EPR-Parametern des Moleküls 2,2,3,4,5,5-hexamethylidaolidin-1-oxyl (HMI) untersucht. Hierfür wurde erneut ausgehend von AIMD-Simulationen die Darstellung des Solvens im Rahmen der EPR-Rechnung variiert, um den Einfluss des Solvens auf EPR-Parameter für die protonierte ($HMIH^+$) und neutrale Spezies (HMI) zu quantifizieren. Obwohl im Rahmen der AIMD-Simulation eine deutliche Veränderung der Solvensumgebung des EPR-aktiven Nitroxids festgestellt werden konnte, hat sich gezeigt, dass zu etwa 75% eine veränderte elektronische Struktur durch das zusätzliche Proton ausschlaggebend für die veränderten EPR-Parameter verantwortlich ist. Lediglich die restlichen 25% konnten der veränderten Solvensumgebung angerechnet werden.

In dem letzten Projekt konnte ein ML-Potential entwickelt werden, dass auf Basis von quantenchemischen EC-RISM-Rechnungen trainiert wurde und in der Lage ist Vorhersagen für die meisten neutralen, organischen Moleküle zu treffen. Die Genauigkeit des Potentials wurde in der Vorhersage von Tautomer-Gleichgewichten getestet, wo ein zusätzlicher Fehler von weniger als 1 kcal/mol im Vergleich zu der direkten Referenzrechnung auftrat. Die Vorhersage auf Basis des ML-Potentials ist dabei jedoch um mehrere Größenordnungen schneller. In der Vorhersage von freien Solvatationsenthalpien zeigte sich ein zusätzlicher Fehler von etwa 1,5 kcal/mol, womit eine vergleichbare Genauigkeit wie mit impliziten Solvatationsmodellen erreicht werden konnte, jedoch erneut mit einem deutlich geringeren Aufwand. Diese vergleichbare Genauigkeit konnte auch in der Optimierung von Geometrien erreicht werden. Es zeigte sich, dass das Training des ML-Potentials an quantenchemischen Rechnungen in Lösung essenziell ist und Gasphasen-Potentiale, wie sie aktuell zu einem großen Teil vertreten sind, bei allen Anwendungen in Lösung deutliche Schwächen aufweisen, auch wenn der Trainingsdatensatz deutlich größer ist.

Durch die enge Verknüpfung zwischen experimentellen, spektroskopischen Messungen und einer Vielzahl von theoretischen Methoden konnten so im Rahmen dieser Arbeit nicht nur weitere strukturelle Veränderungen von Molekülen unter Hochdruckbedingungen oder pH-Änderungen bestimmt werden, ein großer Teil dieser Arbeit bestand auch darin theoretische Methoden im Hinblick auf ihre Genauigkeit und Effizienz zu untersuchen und so Empfehlungen für potenzielle weitere Anwendungen geben zu können.

Abstract

Understanding processes on the molecular level is the central element of chemistry. In addition to experimental techniques such as spectroscopy, this understanding is increasingly being gained through the use of theoretical methods. Historically, quantum chemical calculations or molecular dynamics (MD) simulations have been the methods of choice for gaining deeper insights into chemical processes. In recent years in particular, technological advances in handling large amounts of data have also brought forth artificial intelligence methods, which are now being applied in chemistry as well.

In this work, all these methods were combined to gain a deeper understanding of chemistry at the atomic level. A particular focus was placed on understanding solvation conditions and their changes under extreme conditions such as high hydrostatic pressure. On the experimental side, changes in nuclear magnetic resonance (NMR) spectroscopy parameters under high hydrostatic pressure were investigated. For this purpose, a series of small molecules were measured. These experiments were used to optimize theoretical methods in order to accurately model behavior under high hydrostatic pressure. This involved a combination of quantum chemical calculations and MD simulations. In the quantum chemical calculations, the Embedded Cluster Reference Interaction Site Model (EC-RISM) was used, which is capable of representing high-pressure conditions. This approach was also applied to depict structural changes under high-pressure conditions. By comparing the results with reliable experimental data, the method, including the resulting structural changes, could be validated. It was found that the influence of high pressure on the structure of small molecules and on the conformational equilibrium of two dipeptides is small, but that even these minor structural changes must be taken into account for accurate quantum chemical predictions of NMR parameters. This was attributed to the high structural sensitivity of NMR parameters, which makes them ideal for comparing theoretical methods.

This was further utilized in another project to compare quantum chemical *ab initio* (AI), classical force field, and machine learning (ML)-based MD simulations. Starting from highly accurate AIMD simulations, various methods were tested for representing the solvent within quantum chemical calculations. By comparing with experimental NMR parameters, the validity of the different methods could be assessed. It was found that a combination of explicit solvent molecules near the focused atom together with an EC-RISM background solvation both reduced computational cost and yielded highly accurate predictions. In comparison, deviations in the structural level of the solvation shell and in local solvation could be identified in the classical MD simulation of the same molecules, which correlated with errors in predicting NMR parameters. By contrast, ML-MD provided an accurate representation of the solvation shell but showed weaknesses in solvent representation, which could be mitigated by expanding the training dataset.

In another project of this work, the pH dependence of EPR parameters of the molecule 2,2,3,4,5,5-hexamethylidazolidin-1-oxyl (HMI) was investigated. Starting again from AIMD simulations, the solvent

representation in the EPR calculations was varied to quantify the influence of the solvent on EPR parameters for the protonated ($HMIH^+$) and neutral species (HMI). Although a significant change in the solvent environment of the EPR-active nitroxide could be observed in the AIMD simulations, it was found that about 75% of the change in EPR parameters is due to an altered electronic structure caused by the additional proton. The remaining 25% could be attributed to the changed solvent environment.

In the last project, an ML potential was developed that was trained on quantum chemical EC-RISM calculations and is capable of making predictions for most neutral, organic molecules. The accuracy of the potential was tested for predicting tautomeric equilibria, where an additional error of less than 1 kcal/mol compared to the direct reference calculation was observed. Predictions based on the ML potential, however, are several orders of magnitude faster. For predicting free solvation enthalpies, an additional error of about 1.5 kcal/mol was found, achieving comparable accuracy to implicit solvation models, but again with significantly lower computational cost. Comparable accuracy was also achieved for geometry optimizations. It was shown that training the ML potential on quantum chemical calculations in solution is essential, and that currently predominant gas-phase potentials show significant weaknesses in all applications in solution, even when the training dataset is considerably larger.

Through the close integration of experimental spectroscopic measurements and a wide range of theoretical methods, this work not only determined further structural changes of molecules under high-pressure conditions or pH variations but also devoted a substantial part to evaluating theoretical methods in terms of their accuracy and efficiency, thereby providing recommendations for potential further applications.

1. Introduction

1.1 Outline and aim of this work

Understanding solvation phenomena is the key to understanding most chemical and all biochemical processes that occur in solution and are influenced by the interaction between a solute and the solvent.[1] Solute-solvent interactions include e.g. hydrogen-boding, electrostatic or dispersion interactions. These interactions can then modify the solubility, structure or reactivity of the solute. Understanding biochemical processes or modifying chemical reactions to fulfill a desired target then clearly depends on a deep understanding of solvation phenomena.

While most biochemical processes rely on water as the main solvent even only focusing on biochemical processes occurring in water, the earth houses a wide variety of thermodynamic conditions including high hydrostatic pressure found in the deep sea. Even under these conditions the deep sea is the most extensive habitat on earth, and it supports surprisingly high biodiversity.[2] Understanding the specific biological adaptations to overcome these harsh conditions are of fundamental biological, astrobiological, and biotechnological importance.[3] This includes a deep understanding of the change in solute-solvent interactions at high hydrostatic pressure. Over the years a variety of biophysical techniques have been applied, including calorimetry, spectroscopy, and neutron and X-ray scattering, in conjunction with high-pressure techniques.[3] In addition, over the last years computational methods have gained great importance in the role of investigating chemical and biological phenomena. The advantage of computational methods relies then in the ability to provide insight into solvation phenomena on a molecular level. In particular, computational methods allow for the controlled variation of thermodynamic parameters such as pressure and temperature, enabling the systematic investigation of solute-solvent interactions extending experimental observations. However, the accuracy and computational cost and therefore usefulness of these methods are highly dependent on the specific systems that are investigated and the specific method that was used. Therefore, a benchmark against experimentally validated observations is necessary wherever possible, to establish computational protocols that can be used for properties that are challenging to access experimentally.

In the context of deep sea life, it is experimentally challenging to quantify the structural and solvation response to high hydrostatic pressure and even computational efforts to this effect are sparse, with computational calculations under high pressure often focusing around molecules in solid state[4] or gas phase.[5] However, it is possible to experimentally determine the spectroscopic response to high hydrostatic pressure using NMR spectroscopy which could be used to determine the change of conformational equilibria under high-pressure conditions.[6]

To further explore computational methods to investigate these phenomena and to gain a deeper understanding on the impact of thermodynamic conditions on both structure and solvation the aim of this work is to disentangle the structural and solvation response of NMR parameters. This will be done by combining a variety of theoretical methods including quantum mechanical (QM) calculations, MD simulations and Machine Learning (ML) approaches. The analysis of the structural and solvation response is also extended to EPR parameters at ambient conditions combining AIMD simulations and QM calculations.

With full biological systems such as proteins outside of the current possibilities to model accurately with QM calculations, small organic molecules such as N-methylacetamide (NMA) are used as surrogates for more complex systems such as the protein backbone.[7] The goal is then to not only find quantitatively accurate computation methods but to also vary the approaches to identify computationally efficient calculation methods which have the potential to be used on more complex systems. While varying the theoretical methods used the response to the calculated parameters will be closely monitored to analyze both computational demand and resulting accuracy. Experimental reference data will be chosen accordingly and partially created as part of thesis to benchmark and verify theoretical methods.

In more detail, this work will first focus on the chemical shift as a NMR parameter where a variation of solvent methods using explicit, implicit and hybrid solvent models will determine the solvent-dependence of the chemical shift and identify computationally efficient, but still accurate solvation methods. Here the influence of different methods for MD simulations, namely quantum mechanics driven AIMD simulations, classical force field simulations and machine learning MD simulations will be compared in terms of the structural ensemble they yield. The structures will be compared in terms of their resulting NMR parameters to quantify the response of the chemical shift to structural and solvation changes.

This approach will be extended to high hydrostatic pressure where experimental benchmark data will also be created as part of this work. The response to high pressure will be analyzed both from an experimental point of view and by combining pressure-dependent quantum mechanical calculations and MD simulations. This will once again disentangle the structural and solvation response to high pressure and accurate experimental data will be used to verify theoretical methods. Here, a main focus will be placed on identifying and verifying the appropriate referencing method of the chemical shift, where previously conflicting methods between experiment and calculations have been used.[7][8] With accurate computational methods identified these will be used to analyze the experimentally unknown pressure-dependent conformational equilibrium of two dipeptides. This will be done by fitting the conformational equilibrium to the experimental, population-averaged, chemical shifts of the respective amide protons. This approach relies on accurately calculated chemical shifts of each conformer which will here come from a fitting approach to a training set of amide chemical shifts, measured and calculated as part of this thesis.

In terms of EPR parameters the *g*- and *A*-tensor will be analyzed for a pH-dependent EPR-active spin label. Here the solvation methods used for the EPR calculation will be varied once again to quantitatively disentangle the solvation effect from the change in electronic structure due to the change in protonation. This work extends previously performed calculations[9] on the neutral form of the spin label to the protonated form.

Lastly, a general-purpose machine learning potential (MLP) will be derived from quantum mechanical calculations in solution employing the EC-RISM solvation model. By comparison to direct calculations the accuracy of this approach will be analyzed showing potential use cases in the general workflow of theory-based physicochemical predictions. Comparing the MLP to gas-phase potentials available in literature will point out the necessity to include solvent information within these MLPs if the applications are aimed at processes in solution. In this way the currently predominant method of using explicit solvation derived from gas phase potentials[10] can be changed to an implicit solvation approach further reducing the computational cost. This potential opens up the possibility to perform accurate implicit solvation MD simulations expanding on the previously developed MLPs which rely on a more costly explicit solvation approach. These MD simulations could also again be used for ensemble-based spectroscopic calculations expanding the available methods for accurate and efficient NMR or EPR calculations.

1.2 Pressure-dependent NMR spectroscopy

NMR spectroscopy is one of the most important analytical tools within chemistry to determine the composition, structure and even dynamics of small organic molecules,[11] biomolecules[12] and also solid state materials.[13] With that computational methods used to assist experimental methods e.g. in the determination of relative configurations[14] or the full assignment of spectra to structures have been well established.[15] While the computational methods historically relied on the use of QM calculations, recently ML approaches have been developed to either reduce the computational cost of the structure elucidation or increase the accuracy of computational NMR calculations.[16]

Moving from ambient conditions to extreme conditions an opposing picture emerges in terms of the frequency with which both experimental and theoretical NMR techniques are used. Here, the use of high hydrostatic pressure in the context of NMR spectroscopy often revolves around the investigation of high energy conformations of proteins.[17] The advantage of NMR spectroscopy compared to other techniques lies in the all-atom resolution that is retained at high pressure conditions and allows to investigate local as well as global changes in the structure of a protein. In practice pressures of up to several kilobars can be used to explore the conformational landscape of proteins around the native ground state where the current state of the art uses a high pressure cell that contains the sample within a spectrometer that is kept at ambient conditions.[18] The underlying physical reason behind this technique is that the perturbation due to high pressure does not create new conformations

inaccessible at ambient conditions. It rather reveals low energy excited states that due to a lower partial molar volume become dominant at high-pressure conditions but are still populated to a lesser extent at ambient conditions.[19] This is due to the fact that fast conformational changes on the microsecond scale or less cannot be detected in NMR spectroscopy. The reported structure from an NMR study therefore resembles an average structure of the conformational ensemble. High-pressure NMR can then be used to study different parts of the conformational landscape including open conformations of enzymes,[20][21] partially unfolded states,[22][23] and molten globule (MG) states.[24] To highlight one example from Fuglestad *et al*.[22] they found that the native state of the lactose repressor core domain dimer almost fully vanishes under hydrostatic pressures of up to 3000 bar and partially unfolded states become populated but a monomeric state is still never observed. In this study they used two dimensional TROSY ^{1}H-^{15}N HSQC[25] where they could detect pressure-dependent changes in the chemical shift associated with the conformational change. Reduced intensity at the original chemical shift values at ambient conditions allowed them to quantify the population of the native state at high pressure which becomes unfavorable by 1.4 kcal/mol at 3 kbar. However, by this method that assumes a constant chemical shift for a given conformation chemical shift due to even smaller perturbations such as small changes in bond length might be disregarded and could lead to underestimation of the native state population. In another study Kitahara et al. [26] tried to address this problem by establishing a baseline for chemical shift changes of individual amino acids within the protein backbone by measuring the pressure-dependent chemical shift changes in tetrapeptides of the form Gly–Gly–X–Ala. They modelled chemical shift changes using:

$$\delta_i = a_i + b_i p + c_i p^2 \qquad (1)$$

highlighting the importance of nonlinear changes in the chemical shift with respect to pressure. Within this study they could show large chemical shift changes on the order of -0.5 to 0.5 ppm/GPa for the amide proton of the amino acids. They found that the chemical shift changes can be larger by a factor of ca. 3 in globular proteins highlighting that the specific behavior of the protein in terms of conformational change is still the most dominant part for chemical shift changes, but a certain baseline must be included for the understanding of pressure induced changes. This highlights the importance of the amide protons which will feature heavily in the pressure dependent NMR investigations performed in this work.

To further establish baseline phenomena in high-pressure NMR spectra beyond conformational change chemical shift changes can also be directly correlated with a change in intramolecular hydrogen bond lengths as Li et al. [27] already found in 1998 for the Basic Pancreatic Trypsin Inhibitor (BPTI) and other proteins. They found that the chemical shifts of all amide protons in the backbone increase. Based on known crystal structures at ambient conditions and pressure-dependent chemical shift changes they determined that the intramolecular hydrogen bond length from amide protons to carbonyl oxygens changes by +0.01 Å to −0.11 Å at 2 kbar with the average being −0.020 Å. This corresponds to a shortening of the hydrogen bond of about 1%.

Even though theoretical methods should be ideal to investigate and explain these high-pressure phenomena successful applications of theoretical methods are very limited. This is due to multiple reasons. First, with many experiments revolving around proteins the complexity of protein structures requires a lower level of theory such as molecular mechanics in terms of MD simulations to handle these large systems. On the other hand, the previously mentioned results already highlight that high pressure only introduces very small perturbations to the system in terms of the change in bond length or the energetic difference between conformations. Within NMR spectroscopy the chemical shift as a primary information however is very sensitive to small geometric changes requiring a high precision of theoretical methods to explain small changes. Still, as NMR spectroscopy is a highly reliable and accurate tool it can also act as benchmark for any type of theoretical methods which has been widely used at least at ambient conditions.[28][29][30] Typical errors of high level theoretical calculations which are benchmarked and adjusted to a broad range of molecules can reach an accuracy of ca. 0.2 ppm for ^{1}H [28] or 2 ppm for ^{13}C [31] which can further be improved by the use of neural networks.[32]

In terms of theoretical investigations of high-pressure NMR spectroscopy Frach et al. [7] investigated the protein backbone analogue NMA using EC-RISM[33] and benchmarked their calculations against experimental measurements. They were able to reproduce the experimental trend of an increase in the chemical shift for amide proton, carbon, and nitrogen but overestimated the magnitude of the increase. They noted that all experimentally measured chemical shift changes are in the range of random coil protein highlighting the use of NMA as a suitable model system for the protein backbone that can still be treated with high level quantum chemical calculations. Further pressure dependent NMR calculations using EC-RISM have been performed on the reference compounds DSS and ammonia, and target molecule trimethylamine *N*-oxide (TMAO).[8] Expanding and referencing these previously performed calculations the same molecules will also feature within this work.

1.3 Investigation of Nitroxide EPR spin labels

Moving over from NMR spectroscopy measuring the magnetic effect on a nuclear spin EPR spectroscopy measures the impact of a magnetic field on an unpaired electron spin. Naturally unpaired electrons can occur in the form of reactive oxygen species (ROS), reactive nitrogen species (RNS)[34] or metalloproteins containing V, Mn, Fe, Co, Ni, Cu, Mo or W.[35] Detecting these species or elements can yield insights into reaction mechanisms[36] or the structure of proteins with the class of iron-sulfur proteins as a prominent example, that was first discovered using EPR spectroscopy.[37]

Extending the applications of EPR spectroscopy to natively diamagnetic molecules is done by site-directed spin labeling (SDSL).[38][39] As part of SDSL an EPR active so-called spin-label gets covalently attached to the molecule of interest which can be a protein but also DNA and RNA[40] and even a lipid.[41] Among the most commonly used spin-labels are stable nitroxide radicals usually consisting of a five-membered (pyrrolidine or pyrroline) or six-membered (piperidine ring) with bulky groups on the

ring carbon atoms adjacent to the nitrogen[42] that stabilize the nitroxide radical. Due to their low molecular weight but still structural flexibility that allows for distinct reactivity, nitroxide spin labels can be used for a variety of substrates. As one example the most common spin label MTSL[43] (1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate) reacts specifically with thiol groups and therefore labels surface-exposed cysteine residues of a protein with basically 100% efficiency. With the EPR-spin probe attached to a specific site the structure or dynamic of a protein can be investigated with the large advantage that the remaining protein is EPR-silent and disentanglement of signals, as for example necessary in NMR spectroscopy, is not necessary. Changes in the continuous-wave (cw) EPR signal of the spin-label compared to bulk properties can then give insight into the local pH[44] or electrostatic environment[45] of the protein at the given position of the spin-label. However, to fully understand the dynamic changes of the environment and the influence of the specific spin-label is a difficult task where theoretical investigations can be useful. For example, Owenius et al.[46] studied MTSL in different solvents with varying dielectric constants. They experimentally found a linear dependence of the isotropic (g_{iso}, A_{iso}) and anisotropic (g_{xx}, A_{zz}) EPR parameters. To gain more insight they performed DFT calculations where they varied the dielectric constant and also the number of hydrogen bonds towards the nitroxide group. They found that in the apolar region ($\varepsilon < 25$), the sensitivity of A_{iso} and A_{zz} to ε is large but in the polar region ($\varepsilon > 25$), the sensitivity to ε is small, and the shifts in A_{iso} and A_{zz} are mainly determined by the proticity of the solvent. From theoretical calculations they specifically found that methanol forms on average 1 and water on average 2 hydrogen bonds to the nitroxide. As another example 2,2,3,4,5,5-hexamethylidaolidin-1-oxyl (HMI) was discovered already in 1982 as a pH sensitive spin probe[47] and to this day shows one of the largest changes in A_{iso} upon protonation among nitroxide spin labels[48] with a change of ΔA_{iso} of ca. 3.5 MHz which is comparable to typical peak-to-peak linewidths of X-band EPR spectra of nitroxide in aqueous solution.

Theoretical investigations of EPR spin probes rely mostly on the use of DFT methods due to their relatively low computational demand. Here it is noteworthy that the hyperfine coupling constant A, as it reflects the spin density close to the nucleus, is particularly sensitive to both very tight and diffuse s functions[49] making the choice of an appropriate basis set important. In a benchmark study of 23 nitroxide radicals and a variety of DFT functionals Gromov et al. [50] found that the best performing functional M06[51] together with a quadruple zeta basis set can reach an accuracy in the prediction of A^{iso} of 0.6 G but they also found a number of basis sets with mean absolute error larger than 2 G which they define as a significant deviation from experiments.

In addition to the choice of the level of theory the A^{iso} of a nitroxide radical is also found to be directly influenced by the out of plane angle of the nitroxide group which has to be accounted for accordingly[49] either by vibrational perturbation theory or sampling from MD simulations.[52] Here vibrational perturbation theory is computationally less demanding but numerically less stable.

Going beyond nitroxide radicals Windom et al.[53] found for a benchmark set of 26 small organic molecules with 65 A^{iso} distinct benchmark molecules only high level coupled cluster methods (CCSD(T) or CCSD) were able to reproduce experimental A^{iso} with sufficient accuracy while DFT methods

introduced errors of 5-15 MHz. This is in line with a benchmark study of Sharma et al.[9] of the HMI spin probe where they combined AIMD simulations with high level CCSD calculations of A^{iso} and reached an accuracy of 2 MHz compared to experiments but using DFT calculations the error increased to more than ca. 7 MHz. Extending this study to the protonated form of HMI will be part of this thesis. As the main goal will be the analysis of relative effects between protonated and deprotonated HMI the DFT level of theory is still expected to give reliable results for relative effects.

As another observable within EPR spectroscopy the *g*-tensor as the analogous observable to the chemical shift yields important information about the structure surrounding the EPR active site. In contrast to the chemical shift in NMR spectroscopy the *g* values are often measured in solid state, even for frozen aqueous solutions. Thus, an orientation dependent *g*-tensor is measured, yielding additional information.

With this in mind the *g*-tensor has become a prominent target of computational investigations since the first theoretical framework for the *g*-tensor calculation has been developed in 1997 by Schreckenbach and Ziegler.[54] Here, once again mostly DFT calculations have been employed and the importance of accurately describing the solvent has already been noticed by Rinkevicius et al.[55] who saw an improvement of the calculated *g* tensors of the di-*tert*-butyl nitroxide radical by going from an implicit solvent to explicit solvent molecules. With this in mind, a variety of solvation methods will be used in the calculation of EPR parameters within this thesis with hybrid solvation models based on explicit solvation and additional EC-RISM background solvation expected to be the most accurate.

1.4 Machine Learning Potentials

As the importance of highly accurate computational methods was demonstrated in the previous chapters the computational demand of these methods can be tremendous making them unfeasible for large systems. Using less accurate methods such as molecular mechanics might not yield the sufficient accuracy to gain insight into the systems of interest. To overcome this gap between accuracy and computational demand machine learning models have been developed over the last decades. The goal of these methods is to reach the accuracy of high-level computational methods potentially even coupled cluster accuracy[56][57] but at significantly lower cost, comparable to molecular mechanics methods. This is usually done by training a neural network model to calculated data on the desired level of theory, based on specific so-called features of the training data, i.e. molecular geometries. The goal is that the machine learning model is then able to match the input features to the output data that was previously calculated using the high level method. To exploit the capabilities of the machine learning model it should then be able to interpolate results between known data points and ideally also extrapolate to new regions of chemical space based on the input features used. In practice this can be done by using so-called Behler-Parrinello symmetry functions[58] or derived from these atom environment vectors (AEVs).[10] They model the surrounding of a specific atom based on input

coordinates. A trained neural network is then able to calculate atomic energies based on the symmetry function or AEV which can be summed for all atoms in one molecule. Inter- and extrapolation is possible with these input structures as only the near surrounding of a given atom is considered and needs to be trained upon and new configurations of atoms might be broken down into known fragments.

Another approach that has become popular, called MACE,[59] relies on a different architecture by employing an equivariant message passing neural network (MPNN) to predict atomic energies.

In practice many MLPs are still specialized to reproduce a very narrow chemical space which could be individual elements[60][61] or classes of molecules, such as zeolites.[62] This is due to the fact that a specific MLP can be quickly trained typically requiring on the order of 10^4 structures.[63] The number of structures in the training dataset can be reduced by using techniques such as active learning.[64][65] Here structures are added that are different from already known structures are added on the fly diversifying the training data as much as possible with as a few calculations as possible. Based on the current dataset MD simulations are performed exploring new but still physically reasonable chemical space.

In contrast to specific MLPs trained on a narrow chemical space general purpose MLPs try to cover a large chemical space. To accomplish that they require a diverse training set. For example the ANI-2x dataset contains more than 10^5 distinct molecules with more than 10^7 individual structures.[66] With that the MLP is able to predict gas-phase energies small molecules with an error of ca. 2 kcal/mol compared to DFT calculations covering molecules containing seven different elements and up to 62 atoms. Based on the MACE architecture the models of the MACE-OFF23[67] family are also general purpose MLPs even outperforming ANI-2x in terms of accuracy but due to the multi-layer message-passing have an increased computational cost.

The disadvantage of these general purpose MLPs is that they are then limited by the level of theory that can be used for millions of calculations also limiting the use of solvent models which require more computational time. One example of a dataset containing data for molecules in solution is the 2024 published Aquamarine (AQM) dataset[68] which contains DFT data of ca. 60k conformers of 1.6k small organic molecules calculated using an implicit solvent model and is used to show solvent effects on over 40 global and local physicochemical properties for a broad range of molecules. This dataset is still roughly two orders of magnitude smaller than the ANI-2x dataset of gas-phase data.

One main application of MLPs is in the atomistic simulation of biomolecules.[69] While these ML-MD simulations still are ca. 50 times slower compared to classical force field simulations it is still possible to run simulations on the ns time scale and for systems containing hundreds of thousands of atoms which is currently not feasible using AIMD simulations.[70] With implicit-solvation MLPs, as developed within this thesis, an additional speedup could be achieved by disregarding the explicit solvent molecules accounting for most of the simulated system similar to the popular generalized Born (GB) solvent model.[71][72]

An inherent advantage of ML-MD simulations is that they can by design mimic chemical reactions in terms of breaking and forming chemical bonds which can not be done using classical force fields with a harmonic potential yielding unbreakable bonds. Even specific datasets focused on training reactive MLPs have been created.[73][74] In addition to MD simulations that require forces from the MLPs another application is in the ranking and optimizations of conformers.[75][76] As MLPs are by design analytical expressions of i.e. the potential energy the required gradient for geometry optimizations is readily available. In the special case that the MLP is trained to forces these can also be used for geometry optimizations.[77][78]

2. Theory

2.1 Introduction to quantum chemistry

The central property within quantum mechanics is the wave function which in principle fully describes the corresponding system. With suitable Hermitian operators acting upon this wave function, it is possible to extract all kinds of different observables from it. One example for this is the time independent Schrödinger equation

$$\hat{H}\psi = E\psi \tag{2}$$

which yields the energy E based on the Hamilton operator (Hamiltonian) $\hat{H}$

$$\hat{H} = \hat{T} + \hat{V} \tag{3}$$

which in its general form includes the kinetic energy operator $\hat{T}$ and the potential energy operator $\hat{V}$ that represent the system at hand. A wave function that satisfies this equation is an eigenfunction with respect to the operator and energy values from this equation are eigenvalues and thus from equation (2) distinct, sharp values can be calculated. For wave functions that are not eigenfunction it is still possible to gain expectation values $\langle O \rangle$ with respect to an arbitrary operator $\hat{O}$ as

$$\langle O \rangle = \langle \psi | \hat{O} | \psi \rangle \tag{4}$$

Written in so-called bra-ket or Dirac notation.[79] As solving the Schrödinger equation analytically is only possible for very simple systems such as a hydrogen atom for practical uses a variety of approximation are usually made. Within this thesis all quantum chemical calculations are using the Born-Oppenheimer approximation.[80] This approximation assumes that the nuclei can be treated classically due to the fact that nuclei are much heavier and therefore move much slower than electrons. As a consequence, solving the Schrödinger equation is then only necessary for the electron wave function which has a parametric dependence on the nuclear coordinates. The total energy of the system is then the sum of the electronic energy and the interaction energy between nuclei which can be treated classically.

Even within the Born-Oppenheimer approximation it remains impossible to analytically calculate wave functions of chemical systems such as molecules.

For these systems the basic form for molecular wave functions is based on the linear combination of atomic orbitals (LCAO) technique. Here so-called atomic orbitals φ_m which are based on the

analytically calculated wave functions of the hydrogen atom (or approximations of these wave functions) get combined to form a molecular wave function as:

$$|\psi\rangle = \sum_m c_m |\varphi_m\rangle \tag{5}$$

In this linear combination these atomic orbitals get weighed with coefficients c_m.

In theory it is possible to use the atomic orbitals derived from the exact solution of the Schrödinger equation for the hydrogen atom called Slater type orbitals (STO) that can be written depending on the spherical coordinates r, ϕ and θ as[81]

$$\varphi_{\zeta,n,l,m}(r,\phi,\theta) = N r^{2n-2-1} Y_{l,m}(\phi,\theta) \exp(-\zeta r) \tag{6}$$

containing an angular part with so called spherical harmonics $Y_{l,m}$ and a radial part depending on the distance r with a parameter ζ. However, solving the Schrödinger equation for molecules with STOs is computationally demanding. As a consequence, other types of orbitals have been created. One of these are Gaussian type orbitals (GTOs) which have the form[82]

$$\varphi_{\zeta,n,l,m}(r,\phi,\theta) = N r^{2n-2-1} Y_{l,m}(\phi,\theta) \exp(-\zeta r^2) \tag{7}$$

in internal coordinates or

$$\varphi_{\zeta,i,j,k}(x,y,z) = N x^i y^j z^k \exp(-\zeta r^2) \tag{8}$$

in cartesian coordinates. Resulting integrals based on GTOs are less computationally demanding but they are less accurate than STOs, especially close to the nucleus. To overcome this problem a linear combination of multiple GTOs can be used to resemble STOs. These so-called contracted GTOs are the basis for many popular types of basis sets with different number of GTOs that are contracted different so-called primitive GTOs that get contracted with parameters optimized for to reproduce various observables. This results in wide variety of available basis sets. In this thesis the Pople-type basis sets[83] and the Ahlrichs-type basis sets[84] have been used.

One of the most common methods to calculate wave functions is based on the variational principle (Rayleigh-Ritz method). It states that the energy of the true ground state wave function E_0 is always smaller than for any other wave function:

$$E_0 \leq \langle \psi | \hat{H} | \psi \rangle \tag{9}$$

Based on this criterion it is possible to iteratively calculate the ground state wave function.

With fixed atomic orbitals using the LCAO method the only parts of the wave function that are then varied are the set of coefficients. Using the Lagrange-formalism solving the Schrödinger equation can be rewritten as[85]

$$\mathbf{Hc} = E\mathbf{Sc} \tag{10}$$

with the Hamilton matrix **H** as

$$H_{ij} = \langle \varphi_i | \hat{H} | \varphi_j \rangle \tag{11}$$

and the overlap matrix **S** as

$$S_{ij} = \langle \varphi_i | \varphi_j \rangle \tag{12}$$

In this way solving the Schrödinger equation becomes an algebraic problem of finding the eigenvalues of a matrix.

A further restriction of the molecular wave function is based on the Pauli exclusion principle[86] which requires a multi electron wave function to be antisymmetric under the exchange of the labels of two exact fermions which simplified can be written as

$$\psi(2,1) = -\psi(1,2) \tag{13}$$

with arbitrary coordinates 1 and 2. In addition the electronic wave function also needs to include a spin wave function to differentiate between the two spin states α and β

$$\chi(\boldsymbol{\tau}) = \begin{cases} \phi(\mathbf{r})\alpha(\omega) \\ \phi(\mathbf{r})\beta(\omega) \end{cases} \tag{14}$$

which are present at the same spatial orbital $\phi(\mathbf{r})$.

One efficient way of building molecular wave functions derived from algebraic principles that also inherently follows the Pauli exclusion principle is the so-called Slater determinant[87] which based on the spin orbitals from equation (14) can be written as

$$\psi(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_N) = \frac{1}{\sqrt{N}} \begin{vmatrix} \chi_i(\boldsymbol{\tau}_1) & \chi_j(\boldsymbol{\tau}_1) & \ldots \chi_k(\boldsymbol{\tau}_1) \\ \chi_i(\boldsymbol{\tau}_2) & \chi_j(\boldsymbol{\tau}_2) & \ldots \chi_k(\boldsymbol{\tau}_2) \\ \vdots & \vdots & \vdots \\ \chi_i(\boldsymbol{\tau}_N) & \chi_j(\boldsymbol{\tau}_N) & \ldots \chi_k(\boldsymbol{\tau}_N) \end{vmatrix} \tag{15}$$

2.1.1 The Hartree-Fock and post-Hartree-Fock methods

In one of the most important methods of calculating quantum chemical properties, the Hartree-Fock (HF) method,[88][89] only one Slater determinant is used. In this way only the HF ground state is used, disregarding all excited states. In another approximation within the HF method, the many-body problem is reduced to an effective single-body problem using the Fock operator $\hat{F}$ acting on electron i, written in atomic units as

$$\hat{F}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{i,A}} + \sum_{j=1}^{N/2}\left[2\hat{J}_j(i) - \hat{K}_j(i)\right] \tag{16}$$

where the first two terms represent a one-electron Hamiltonian with the Nabla operator ∇ extracting the kinetic energy. The number of protons Z_A of each nucleus A are part of the Coulomb interaction between electron and nucleus under the Born-Oppenheimer approximation. The last term describes the electron-electron interaction in a so-called mean-field approach with the Coulomb operator $\hat{J}$ and the exchange operator $\hat{K}$. The Hamiltonian of a system can then be approximated by the sum of all Fock operators. Combining all the previously discussed the Schrödinger equation can be solved iteratively using the Fock operator until convergence which is called a self-consistent field (SCF) method.

One limitation of the HF method is the fact that it is using just one Slater determinant. With one Slater determinant the influence of one electron on the wave function of another electron cannot be calculated. This is the so-called electronic correlation and the missing energy contribution from the HF limit compared to the exact solution of the Schrödinger equation is called the correlation energy.[90] With the disregard of the electronic correlation the London dispersion interactions[91] as a result of long range electron correlation are also disregarded In terms of Slater determinants the exact solution of the Schrödinger equation would require a linear combination of all possible Slater determinants which is called the full configuration interaction (FCI). As calculations using the FCI are very time consuming approximations to the FCI but extending on the HF method to include electronic correlation have been made. These methods are called post-HF methods. One of these is the Møller-Plesset-perturbation theory of n-th order (MPn).[92] Here the unperturbed Hamiltonian $\hat{H}_0$ is extended by a perturbation term $\lambda\hat{V}$:

$$\hat{H} = \hat{H}_0 + \lambda\hat{V} \tag{17}$$

Based on Rayleigh-Schrödinger perturbation theory the perturbation is expressed as a power series in terms of λ where the first term resembles a HF calculation. The perturbed Hamilton operator is then applied on an unperturbed wave function e.g. from a previous HF calculation. As an additive post-HF Coupled Cluster theory should briefly be mentioned. Here Slater determinants corresponding to excited states are included up to a certain level. For example CCSD includes the excitations of a single

orbital and two orbitals (doubles) and in CCSD(T) triple excitations are estimated by a non-iterative calculation using arguments from many body perturbation theory is considered the “gold standard” of ab-initio methods for small molecules almost reaching FCI quality.[93]

2.1.2 Quantum mechanical density functional theory

Apart from the HF and post-HF methods described in the previous chapter one of the most used methods of computational chemistry is the density functional theory (DFT). The basis of DFT is the Hohenberg-Kohn theorem.[94] It states that the ground-state of a system can uniquely be represented by one electron density distribution. As a result the same variational principle as described before can be applied. The only necessary part is an accurate calculation of an energy based on the density distribution which within DFT is done by the energy functional

$$E\left[\rho(\mathbf{r})\right]=T\left[\rho(\mathbf{r})\right]+J\left[\rho(\mathbf{r})\right]+\int V_o(\mathbf{r})\rho(\mathbf{r})d\mathbf{r}+E_{xc}\left[\rho(\mathbf{r})\right] \tag{18}$$

which depends on the kinetic energy functional $T\left[\rho(\mathbf{r})\right]$, the Coulomb functional $J\left[\rho(\mathbf{r})\right]$, an external potential $V_o(\mathbf{r})$ and exchange-correlation functional $E_{xc}\left[\rho(\mathbf{r})\right]$. This once again is a mean-field approach to describe the influence of all other nuclei and electrons. Nowadays so-called Kohn-Sham DFT[95] is mostly used compared to the original Thomas-Fermi model[96][97] which was the first model to describe the system based on electron densities. The advantage of Kohn-Sham DFT is that the kinetic energy functional can directly be written as:

$$T\left[\rho(\mathbf{r})\right]=\sum_{i=1}^{N}\int\varphi_i(\mathbf{r})\left(-\frac{1}{2}\right)\nabla^2\varphi_i(\mathbf{r})d\mathbf{r} \tag{19}$$

using atomic units. Here $\varphi_i(\mathbf{r})$ are so-called Kohn-Sham orbitals which can be the same orbitals as described before for wave-function based methods with the electron density resulting from the square of the function. For a DFT calculation the exchange-correlation functional $E_{xc}\left[\rho(\mathbf{r})\right]$ remains difficult to calculate and approximations have to be made. In a first approximation, the local density approximation (LDA), the density distribution is approximated by local point charges for which the exchange functional based on the homogenous electron gas is analytically known as[98]

$$E_x^{LDA}\left[\rho(\mathbf{r})\right]=-\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int\rho(\mathbf{r})^{4/3}d\mathbf{r} \tag{20}$$

and the correlation functional can be well approximated by analytical expressions at the high- or low-density limit.[99][100] A more detailed approximation of exchange-correlation functional is the generalized gradient approximation (GGA).[101][102] Here the gradient of the density distribution is included. In a subsequent development, meta-GGA (mGGA) functionals[51][103] include second derivates of the density distribution corresponding to the Kohn-Sham kinetic energy density. Lastly hybrid functionals can combine HF exchange functionals with LDA or GGA functionals. Here parameters which describe the combination of functionals are usually fitted to experimental data such as atomization energies, ionization potentials, proton affinities or total atomic energies.[104]

With DFT still being a mean field approach the same problem as for HF occurs in the disregard of dispersion interactions which are crucial for the accurate description of intermolecular, noncovalent interactions. To overcome this problem a large number of dispersion corrections, especially for DFT calculations, have been developed. The most prominent correction schemes can be classified into three groups.[105] nonlocal, density-based corrections, like vdW-DF2[106], modify the electronic potential to account for dispersion. Semiclassical C_6-based corrections, like the D3[107] and D4[108] correction by Grimme, apply a correction to the total energy based on an explicit calculation of the nonlocal correlation energy. Lastly, atom centered one-electron effective potentials can be added to correct the external potential otherwise generated by the nuclei.[109] All of these methods however require the use of empirical correction terms which are fitted to reproduce high level post-HF data.

Within this thesis the OLYP[110][111] GGA functional, the r^{2}SCAN[103] mGGA functional which, by design includes the D4 correction,[105] and the B3LYP-[111][102] and revPBE0[112][113] hybrid functionals are extensively used emphasizing the broad range of applications for DFT functionals with different theoretical complexity and computational effort.

2.2 Molecular dynamics simulation

With MD simulations as one of the most widespread techniques of computational chemistry the following part should give a short overview of the most fundamental aspects for this thesis. Numerous review articles over the years [114][115][116][117] are referred to for a deeper insight into the applications of MD simulations. Classical MD simulations allow the time-resolved simulation of systems that are too complex for quantum mechanical treatment with whole genes containing 10^9 atoms successfully simulated.[118] Within classical MD simulation the movement of atoms follows Newtons second law:

$$\mathbf{F}_i = -\nabla \mathbf{V} = m_i \mathbf{a}_i \tag{21}$$

Where the force $\mathbf{F}$ acting on each atom i is calculated by the gradient of the potential $\mathbf{V}$ and related to the acceleration $\mathbf{a}$ by the respective mass m. Even though the acceleration can be used to calculate the time resolved position of atoms, the so-called trajectory, this can not be done analytically. Instead

an iterative approach is used which is called the Verlet algorithm.[119] The combination of a Taylor series in positive $t+\Delta t$ and negative time $t-\Delta t$ up to the second derivative yields:

$$\mathbf{r}_i(t+\Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \mathbf{a}_i(t)\Delta t^2 + O(\Delta t^4) \tag{22}$$

With the advantage that the remaining error is on the order $O(\Delta t^4)$ instead of the expected $O(\Delta t^3)$ for a Taylor series up to the second term. The potential that is used in equation (21) is derived from so called force fields, e.g. the AMBER force fields, which have the following form:[120][121]

$$\begin{aligned} V(\mathbf{r}) = & \sum_{\text{bonds}} K_b (r - r_0)^2 + \sum_{\text{angles}} K_\vartheta (\vartheta - \vartheta_0)^2 \\ & + \sum_{\text{dihedrals}} \frac{V_n}{2}\left(1 + \cos(n\varphi - \gamma)\right) \\ & + \sum_{\text{non-bonded}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon_0 r_{ij}} \end{aligned} \tag{23}$$

Bonds and angles are modelled by a harmonic potential with the force constant K_b around equilibrium distance r_0 and K_ϑ around equilibrium angle ϑ_0. For dihedral angles a sum of individual potentials modelled by cosines is used where the amplitude is modelled by V_n and the phase can be modulated by the parameter γ to model different chemical configurations. For most systems a sum of dihedral potentials up to n=3 is already enough.[85]

Non-bonded interactions are modelled by a Coulomb potential for electrostatic interactions with partial charges q_i and q_j and Lennard-Jones potentials, usually a 12-6 potential, where the parameter σ_{ij} describes the range of the interaction and ε_{ij} models the strength of the interaction.

To minimize the number of pairwise interactions that need to be calculated a cutoff is introduced and interactions between atoms further apart than this cutoff are neglected. The cutoff potential is usually artificially switched to zero instead of a hard cutoff ensuring a continuously differentiable potential energy surface.

To minimize the necessary number of atoms for a simulation periodic boundary conditions (PBC) are used to represent an infinite system by replicating a smaller portion of the system called the unit cell. This allows particles that leave one side of the simulation box to reappear on the opposite side, simulating a continuous simulation box. In theory this could lead to overcounting interaction pairs which is overcome by the minimum image convention that needs to be fulfilled. It states that a given atom interacts only with the nearest replica of any other atom and limits the smallest box size usable for a given cutoff.

By default, based on the Newtonian equations of motion a MD simulation would sample a microcanonical ensemble (*NVE*) with constant number of particles *N*, constant volume *V*, based on the

defined box size of the unit cell and constant total energy *E*. To model different thermodynamic ensembles, matching experimental conditions, it is possible to run MD simulations at constant temperature or pressure.

For constant temperature simulations a thermostat is necessary. Here a variety of thermostats have been developed over the years. They can be divided into two categories.[122] First, velocity randomizing algorithms of which examples are the Andersen[123] and the stochastic dynamics thermostat.[124] The idea is that velocities are updated with some degree of randomness but still following the Maxwell-Boltzmann distribution in the case of the Andersen thermostat or following the Langevin equation of motion in case of the stochastic dynamics thermostat. In contrast velocity scaling algorithms like the Berendsen,[125] V-rescale[126] and Nosé-Hoover[127][128] thermostats take the current velocities of all atoms and rescale them to match the target temperature. Here a coupling constant determines how fast velocities are rescaled to match the temperature. The choice of an adequate thermostat based on the goals of the simulations is important as they each come with their own benefits and problems. For example, the Berendsen thermostat does not yield the correct thermodynamic ensemble but it is highly efficient and can be used for equilibrations.[122] In contrast, the Nosé-Hoover thermostat does yield the correct ensemble but might introduce oscillations for systems far away from equilibrium.

To perform MD simulations at constant pressure a barostat has to be used. Again, multiple algorithms have been developed to keep the system at a given pressure. The Berendsen barostat[125], like the Berendsen thermostat, uses a first-order kinetic relaxation of the pressure towards a reference pressure with a chosen coupling constant. The Parrinello-Rahman barostat[129] is similar to the Nosé-Hoover thermostat as it couples the system to an external, virtual bath and adds a friction parameter to the equations of motion. This could again yield unphysical oscillations if the system is far from equilibrium but the Parrinello-Rahman barostat yields the correct *NpT* ensemble. Depending on the goal of the MD simulation thermostats and barostats should be chosen accordingly.

In contrast to classical MD simulations that rely on parameterized force fields to determine interatomic forces ab-initio MD simulations (AIMD) use QM electronic structure calculations at each time step to determine interatomic forces. Here usually DFT in form of a GGA- or sometimes hybrid functional is employed as the electronic structure method, but in theory every electronic structure method presented earlier can be employed for AIMD simulations with the computational cost as the main limiting factor.[130] Especially, if using DFT for AIMD simulations which is the currently preferred choice for the level of theory, due to its low computational cost, dispersion corrections need to be applied to accurately simulate the long range, intermolecular interactions.

In terms of methods of the AIMD simulation Born-Oppenheimer (BO) MD[131] and Car-Parrinello (CP) MD[132] both assume the Born-Oppenheimer approximation separating electronic and nuclear coordinates and the so-called adiabatic approximation[133] stating that the electrons of the system follow nuclear motion and their motion can be integrated out leading to just one Born-Oppenheimer potential energy surface for the nuclei.[134] In BOMD the quantum mechanically calculated forces are used to solve the classical Newtonian equations of motion while in CPMD the equations of motion are

modified by coupling the nuclear and electronic dynamics within a single extended Lagrangian. With that the forces acting on the nuclei depend on the instantaneous electronic configuration. This is in contrast to BOMD, where for each step a full SCF cycle is required to force the electrons to adiabatically follow the nuclei. [135] However, the required time step for CPMD is usually significantly shorter than in BOMD as the change in electronic configurations directly impacts the equations of motion. Within this thesis only reference data from BOMD simulations is used.

Due to the high computational demand of AIMD the systems and timescales treated with this method are much smaller compared to classical force field MD. The typical system size is on the order of tens to a few hundreds of atoms simulated over a time scale of picoseconds, both depending on the level of theory used. With these limitations, however it is still possible to simulate a solvated small molecule with sub-molar concentration and structural ensembles generated by AIMD are considered the gold standard. A recent review[135] highlights the successful application of AIMD simulations for the ensemble generation for subsequent NMR calculations. This approach is also followed within this thesis with AIMD-based structural ensembles are both used for NMR predictions with the highest possible accuracy and to benchmark FFMD simulations in terms of structural parameters.

2.3 Introduction to the ANI-type Machine Learning Potentials

Another recent development to bridge the gap between the costly but accurate AIMD simulation and the cheap but sometimes inaccurate FFMD simulation is the use of MLPs for simulations. Starting with initial work from Behler and Parrinello [58] the field of MLPs gained increasing interest over the years. In the field of general purpose or universal MLPs the ANAKIN-ME (ANI) method developed by Roitberg et al. and the specific models within are among the most used ones covering a large chemical space.

Potentials like the ANI potentials can be used for MD simulations if they are trained to calculate forces for all used atoms.[66] The calculation speed is roughly 50 times slower than FFMD but still more than two orders of magnitudes faster than AIMD simulations yielding speeds of multiple nanoseconds per day on one GPU for systems with ca. 500 atoms as shown in the following chapters. Limitations might come into place based on the training data that is used to train these models.

However, in the history of ANI potentials the goal to create as broad training data as possible was at the center to create a general-purpose MLP. Starting from ANI-1[10] with more than 17 million individual conformations from more than 56000 molecules containing the elements H, C, N and O active learning reduced the number of conformations for the ANI-1x model to roughly 25 % of ANI-1.[136] Further steps in generalizing the use of the ANI potentials was made with ANI-2x where the elements F, S and Cl were added in the training. The ANI-2x model[66] and dataset containing almost 10 million structures is the latest ANI model. ANI-1, ANI-1x and ANI-2x were all calculated on the ωB97X-6-31G*[137] level of theory, however another model, called ANI-1ccx[138] was trained to high

level CCSD(T)/CBS calculations, which were done for a reduced set of 500 k structures containing the elements H, C, N and O.

Overall, these models are able to reach accuracies close to the DFT level of theory that they were trained on for a variety of tests. For example the ANI-2x model was able to reach an RMSE of 0.59 kcal/mol[66] compared to CCSD(T)/CBS calculations on the Genentech Torsion Benchmark.[139] This is only slightly worse than the RMSE of 0.51 kcal/mol for the direct DFT calculation and outperforms classical MM calculations. In a similar vein, both ANI-2x and ANI-1ccx were even able to outperform direct B3LYP calculations of the conformational energy profile of organic molecules.[140]

To allow for universal use of the MLP the model itself must include a large degree of flexibility while still making use of the training data that is available. For the ANI model structural information is encoded in Atomic Environment Vectors (AEVs). These expand on the symmetry functions developed by Behler [141] and give a vector that describes the environment of one atom. The AEVs are divided into radial and angular parts which are further subdivided by elements and shifted by

$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot \cos\left(\frac{\pi R_{ij}}{R_c}\right) + 0.5 & \text{for } R_{ij} \leq R_c \\ 0.0 & \text{for } R_{ij} > R_c \end{cases} \tag{24}$$

and

$$G_m^R = \sum_{j \neq i}^{\text{all atoms}} \exp\left(-\eta (R_{ij} - R_s)\right)^2 f_c(R_{ij}) \tag{25}$$

Here η changes the width of the Gaussian distribution and R_s shifts the center of the peaks and f_c approximates the local environment up to a cutoff which is ca. 5 Å among ANI models.

To illustrate how AEVs are used within an ANI model Figure 1 shows how atomic energies and molecular energies are calculated. Here the AEVs $\mathbf{G}_i$ are calculated from the coordinates and taken as the input for a neural network. The output is an atomic energy. For a molecule this individual step is repeated for all atoms and the sum of atomic energies yields the molecular energy.

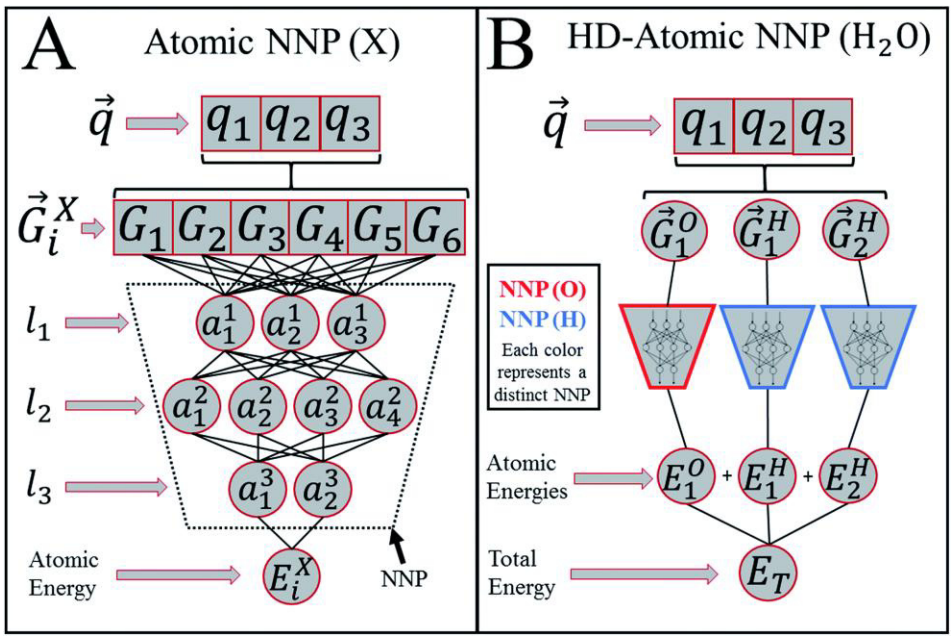

Figure 1: Illustration of an atomic neural network potential (NNP) (**A**) and for H_2O (**B**). The picture was taken from Ref.[10]

As atomic and molecular energies can yield a large range of values and the desired prediction accuracy is in the range of a few kcal/mol atomic energies can be normalized by so-called ground state atomic energies (GSAEs) which represent the atomic energy of an individual atom at the training level of theory. The output of an ANI-type potential can then be interpreted as the atomization energy of the molecule.

To further illustrate the AEVs as the input for all ANI-type potentials Figure 2 shows the example of formic acid in two conformations. It shows the extra parameters introduced by using many shifted symmetry functions. With radial and angular terms different conformations of the same molecule including enantiomers can be described by widely different AEVs.

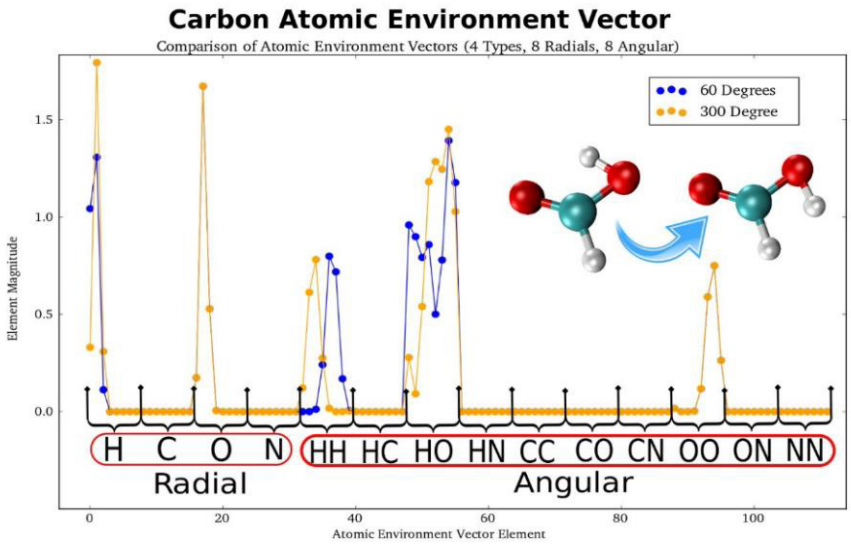

Figure 2: Atomic environment vector (AEV) for the carbon atom in two different conformations of formic acid. The picture was taken from Ref.[10]

With this representation by AEVs ANI models make use of many input parameters which allow for the general use of a singular trained model. With a relatively short cutoff of ca. 5 Å ANI models focus on the very local environment of the atom. This allows the training dataset to consist of small molecules as more than 5 Å surrounding the central atoms will not be included. Here an active learning strategy with an increasing number of atoms in the molecules that are tested was employed and reduced the size of the necessary dataset.[136] For ANI-2x the number of structures in the dataset sorted by the number of atoms is shown in Figure 3. A Kernel density estimate (KDE)[142] is shown as a continuous interpolation of the discrete data points corresponding to the whole number of atoms.

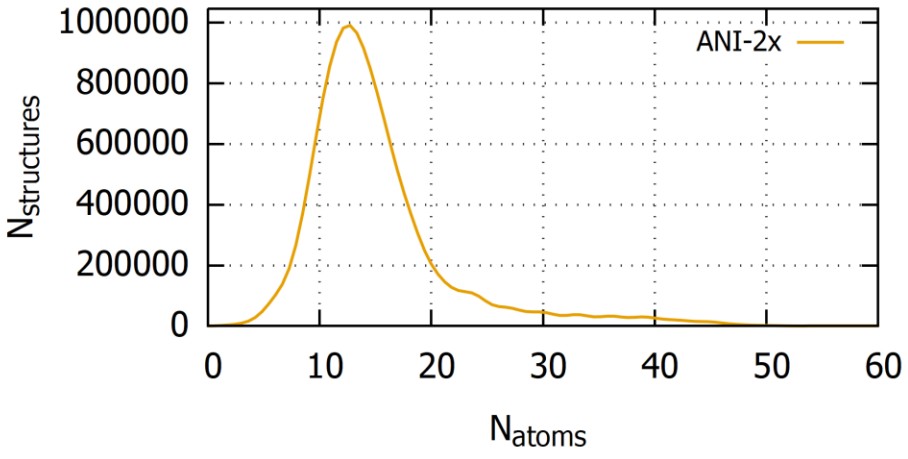

Figure 3: Kernel density estimate (KDE) of the number of structures in the ANI-2x dataset[66] grouped by the number of atoms.

Figure 3 shows that most structures consist of less than 20 atoms (including hydrogen) showing that small molecules are sufficient to train an ANI model due to cutoff in the AEVs and further active learning, that was performed in the creation of this dataset only yielded minor amounts of larger molecules, that were deemed necessary to be included.

2.4 Solvent models

As most processes are performed in solution, every theoretical representation of these processes relies on the accurate description of the interactions between solute and solvent. Figure 4 shows a systematic overview of the solvent models used in this work, which will be more thoroughly described afterwards. The most direct approach is the representation of explicit water molecules (**E**). Here no formal differences are made between solute and solvent. This approach is computationally most demanding as it requires sampling of solvent while treating the solvent on the same level of theory as the solute. Therefore, it is most common only in MD simulations where the multitude of interactions can be calculated in reasonable time.

Within this work it is also used within QM calculations for a limited number of solvent molecules and based on MD simulations to sample the specific solvent structures. In contrast implicit solvation models describe the solvent as a dielectric continuum which surrounds the solute cavity (**A**). In between these models the EC-RISM (**B**) retains some of the structural features of the solvent by calculating the site-specific 3D solvent density. With that regions of high local solvent density and low local solvent density can be seen. These are present independently for all modelled sites of the solvent, which usually correspond to the atoms of the solvent. For the case of water and other protic solvent this imitates directed interactions like hydrogen bonds which cannot be modelled by an implicit solvent. This is also the case when explicit water is added as shown in **E** where high oxygen density is visible next to the explicit water molecule extending a hydrogen bond network. A combination of implicit or EC-RISM solvent with explicit solvent (**D**,**E**) will also be used widely within this thesis.

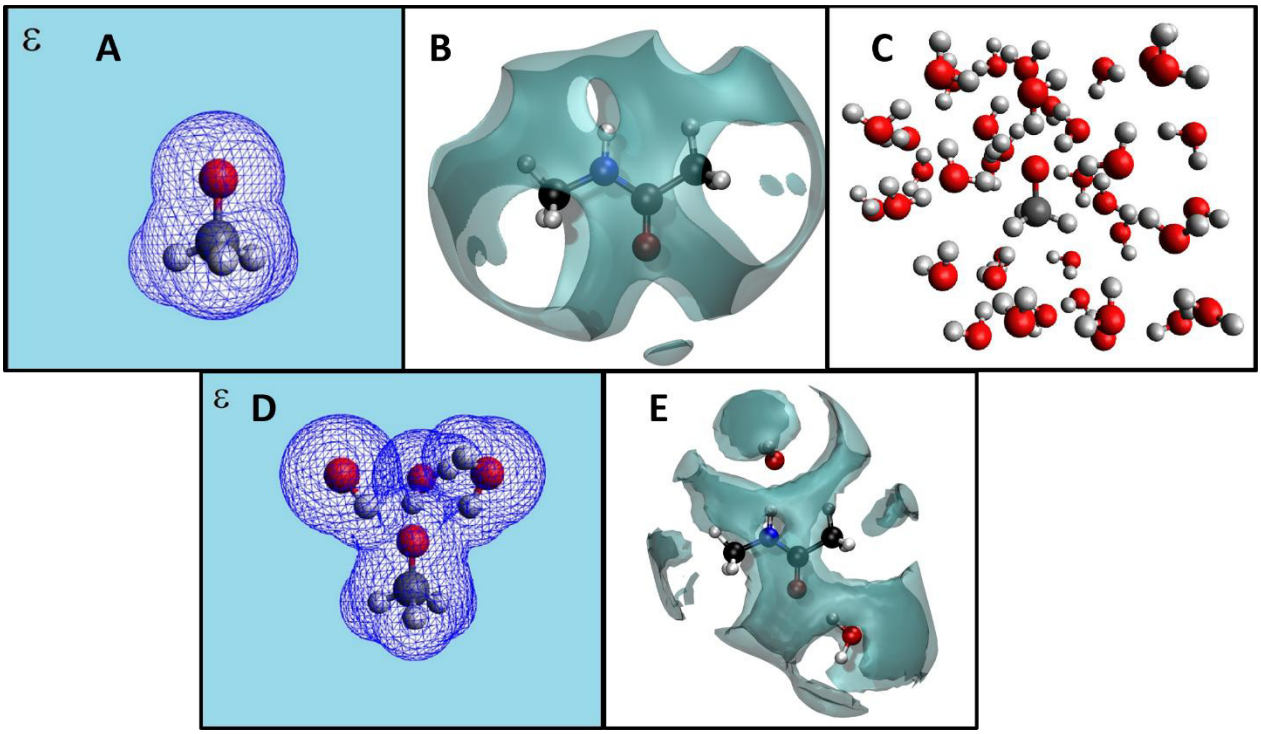

Figure 4: Schematic representation of a continuum solvent (**A**) and EC-RISM solvent (**B**), together with hybrid solvent systems containing a continuous background (**D**) and EC-RISM background (**E**). The pure explicit solvent is shown in **C.** For EC-RISM only an isosurface of the oxygen distribution is shown for clarity. **A** and **C** and **D** taken from Pliego *et al*. [143]

The theoretical details of these models will be outlined in the following parts.

2.4.1 Implicit solvent models

To reduce the computing time compared to an explicit solvation system the general approach of implicit solvation models is to split the overall Hamiltonian $\hat{H}^{\mathrm{FR}}(\mathbf{f},\mathbf{r})$ describing the whole solvated system into a focused model F which represents the solute and the remainder R which represents the solvent. [144]

$$\hat{H}^{\mathrm{FR}}(\mathbf{f},\mathbf{r}) = \hat{H}^{\mathrm{F}}(\mathbf{f}) + \hat{H}^{\mathrm{R}}(\mathbf{r}) + \hat{H}^{\mathrm{int}}(\mathbf{f},\mathbf{r}) \tag{26}$$

Within the framework of implicit (continuum) solvation models $\hat{H}^{\mathrm{R}}(\mathbf{r})$ is reduced to zero as the degrees of freedom of the solvent are not accounted for explicitly which yields an effective Hamiltonian to describe the solute.

$$\hat{H}^{\mathrm{FR}}_{\mathrm{eff}}(\mathbf{f},\mathbf{r}) = \hat{H}^{\mathrm{F}}(\mathbf{f}) + \hat{H}^{\mathrm{int}}(\mathbf{f},\mathbf{r}) \tag{27}$$

During a QM calculation the interaction of the solute with the solvents induces charges in the dielectric medium, which in turn acts back on the molecule. The interaction with the solvent model must thus be calculated by an iterative procedure called Self-Consistent Reaction Field (SCRF).[85]

For QM calculations a variety of models have been successfully developed based on this approach with widespread applications. Some of the best known models are the polarizable continuum model[145] (PCM) or the conductor-like screening model[146][147] (COSMO). The basis to describe the solvent purely by electrostatic interaction of a dieletric continuum is the classical Poisson equation

$$-\vec{\nabla}\times\left[\varepsilon(\mathbf{r})\vec{\nabla}V(\mathbf{r})\right]=4\pi\rho_M(\mathbf{r}) \tag{28}$$

here $\varepsilon(\mathbf{r})$ is a position dependent dielectric constant, V is the sum of the potential generated by the charge distribution ρ_M of the solute and the potential generated by the polarization of the dielectric continuum. The cavity C within these models is defined by a permittivity of 1 inside the cavity and ε outside of the cavity constant outside of the cavity. This simplifies equation (30) and yields

$$\begin{aligned}&-\vec{\nabla}^2V(\mathbf{r})=4\pi\rho_M(\mathbf{r}) \text{ within } C\\&-\varepsilon\vec{\nabla}^2V(\mathbf{r})=0 \text{ outside } C\end{aligned} \tag{29}$$

with the assumption that all solute charges are within the cavity. At the surface of the cavity boundary conditions must be enforced that prohibit jumps in the potential $V(\mathbf{r})$ and it's first derivative. One approach that enforces these boundary conditions and is an exact representation of the potential generated by the polarization of the dielectric continuum is the apparent surface charge (ASC) method which defines a potential V_σ

$$V_\sigma(\mathbf{r})=\int_\Gamma\frac{\sigma(\mathbf{s})}{|\mathbf{r}-\mathbf{s}|}\mathrm{d}^2s \tag{30}$$

Based on the surface charges $\sigma(\mathbf{s})$ at the surface Γ of the cavity. Different implicit solvent models differ in the way the cavity is determined and surface charges are calculated. For the PCM (also called dielectric or D-PCM) surface charges are calculated as:[144]

$$\sigma(\mathbf{s})=\frac{\varepsilon-1}{4\pi\varepsilon}\frac{\partial}{\partial\mathbf{n}}(V_M+V_\sigma)_{\text{in}} \tag{31}$$

where the gradient is computed on the internal (in) part of the surface with the vector **n** being the gradient vector perpendicular to the surface and pointing outwards. Other formalisms based on the original PCM include the integral equation formalism[148] (IEFPCM) and the polarizable conductor model[149] (CPCM). Especially the CPCM is also closely related to the COSMO[146] or COSMO-RS[147]

solvation model where the dielectric constant is initially set to infinity corresponding to a conductor. To still recover a finite ε corresponding to real solvents surface charges are calculated using

$$\sigma(\mathbf{s}) = f(\varepsilon)\sigma^*(\mathbf{s}) \tag{32}$$

where the ideal unscreened charge density $\sigma^*(\mathbf{s})$, which corresponds to $\varepsilon = \infty$, is scaled by an empirical function $f(\varepsilon)$ of the type

$$f(\varepsilon) = \frac{\varepsilon - 1}{\varepsilon + k} \tag{33}$$

where *k* is fitted for the respective solvent within the COSMO model and is set to 0 for the CPCM formalism. Within this thesis either the IEFPCM or the CPCM were used.

In 2008 Cammi et al. [150] developed the XP-PCM which can be used to perform QM calculation at extreme pressures. The pressure is monitored by

$$p = -\left(\frac{\partial G_{\text{el-rep}}}{\partial V}\right) \tag{34}$$

where $G_{\text{el-rep}}$ is the basic energetic quantity of the PCM model entailing the PCM electronic energy with Pauli-repulsion . In essence the influence of high pressure modifies the cavity volume *V* of the solute within the PCM framework while maintaining the cavity shape. In addition the dielectric constant and density of the solvent are adjusted to match the behavior at high pressure.[151] So far, XP-PCM has been used to make predictions of reaction equilibria of 1,3-cyclohexadiene dimerizations[152] or vibrational spectroscopic data of P_4S_3 crystals[153]. Recently it was used to investigate molecular and transition-state geometries.[154] It is noteworthy that high-pressure calculations in an aqueous solution have not yet been performed with XP-PCM.

As an additional computational method similar to XP-PCM the Gaussians On Surface Tesserae Simulate HYdrostatic Pressure (GOSTSHYP) approach was developed in 2021.[155] Here, a set of Gaussian potentials is distributed evenly on the van der Waals surface of the investigated system. This leads to the compression of the electron density and the atomic scaffold.

Both XP-PCM and GOSTSHYP were used in 2024[5] to analyze the bond lengths of a set of diatomic molecules in gas phase. Both models agree on a decrease of the bond lengths and found an increase of the dipole moment of water which was already found in earlier theoretical investigations.[156]

2.4.2 Statistical-mechanical solvation models

In addition to the previously described implicit and explicit solvation models, statistical-mechanical solvation models have been developed over the years. As shown previously in Figure 4 and in contrast to implicit solvation models they retain some solvent structure and can represent directional solvent interactions such as hydrogen bonds but they do not require the exhaustive sampling of solvent structures in contrast to explicit solvation models. These directional solvent interactions are especially important for water as a solvent which makes statistical-mechanical solvation models especially interesting to describe all biochemical processes.

One example of a statistical-mechanical solvent model is the integral equation theory. It aims to represent the solvent distribution which could be the result of exhaustive sampling of an explicit solvation model but at a much cheaper computational cost. The basis of this theory is derived from classical density functional theory (cDFT) and states that the free energy of a liquid is uniquely minimized by one particle density.[157] The grand potential is then defined as a functional of the particle density

$$\Omega[\rho] = \int d\mathbf{r}\rho(\mathbf{r})V_{\text{ext}}(\mathbf{r}) + F[\rho] - \mu\int dr \tag{35}$$

which allows for a variational approach to determine the ground state density

$$\left.\frac{\delta\Omega[\rho]}{\delta\rho[r]}\right|_{\rho_0} = 0\,. \tag{36}$$

As the main thermodynamic quantity of interest the Helmholtz free energy is, similar to electronic DFT mentioned before, uniquely connected to the density by a functional $A[\rho]$. This functional of the particle density can further be split into an ideal (id) part and the remaining excess free energy $A^{\text{ex}}[\rho]$ [158]

$$A[\rho] = A^{\text{id}} + A^{\text{ex}}[\rho]\,. \tag{37}$$

The excess free energy can then be expanded up to the second term as

$$A^{\text{ex}}[\rho] = A^{\text{ex}}[\rho_0] - \beta^{-1}\left(\int \Delta\rho(\mathbf{r})\frac{\delta A^{\text{ex}}[\rho_0]}{\delta\rho(\mathbf{r})}\mathrm{d}\mathbf{r} + \int_0^1 \mathrm{d}\lambda\left(1-\lambda\right)\iint \Delta\rho(\mathbf{r})\Delta\rho(\mathbf{r}')c\left(\mathbf{r},\mathbf{r}';\lambda\right)\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'\right) \tag{38}$$

where $A^{ex}[\rho_0]$ is the unknown excess free energy of the undisturbed reference system and $c(\mathbf{r},\mathbf{r}')$ is the second order correlation function. The integral over the coupling parameter λ is a thermodynamic integration to scale between no solute-solvent interaction ($\lambda = 0$) and full interaction ($\lambda = 1$).

With an approach to determining the particle density briefly outlined and the connection to thermodynamic properties given, an appropriate solvent model should be able to not only describe solvation of spherical particles but common solvents in general.

2.4.3 Reference Interaction Site Model

The statistical-mechanical solvent model used within this thesis is the Reference Interaction Site Model (RISM). Further derivations for the excess free energy and the particle density will be performed in the context of RISM.

The theoretical basis of this model goes back to the Ornstein-Zernike equation[159]

$$h^{(2)}(\mathbf{r},\mathbf{r}') = c^{(2)}(\mathbf{r},\mathbf{r}') + \int c^{(2)}(\mathbf{r}',\mathbf{r}'')\rho^{(1)}(\mathbf{r}'')h^{(2)}(\mathbf{r}',\mathbf{r}'')d\mathbf{r}'' \tag{39}$$

connecting the total correlation function $h^{(2)}$ of two molecules with the direct correlation function $c^{(2)}$ and the indirect influence by a third molecule given by the integral over $\mathbf{r}''$. It can be derived from the grand partition function by the following functional differentiations in the theory of cDFT.[160]

The grand partition function is defined as[158]

$$\Xi[\psi] = \sum_{N=0}^{\infty} \frac{1}{N!} \int \ldots \int \exp(-\beta V_N) \left(\prod_{i=1}^{N} \frac{1}{\Lambda^3} \int \exp(-\beta \psi(\mathbf{r}_i)) \right) d\mathbf{r}_1 \ldots \mathbf{r}_N \tag{40}$$

with $\psi(\mathbf{r}_i)$ called the intrinsic chemical potential. From the grand partition function the grand potential functional is given by

$$\Omega[\psi] = -k_B T \ln \Xi[\psi]. \tag{41}$$

The average density can then be obtained by the functional differentiation

$$\rho(\mathbf{r}) = \frac{\delta \ln \Xi[\psi]}{\delta \psi(\mathbf{r})}. \tag{42}$$

Further differentiation yields:

$$\frac{\delta \rho(\mathbf{r})}{\delta \psi(\mathbf{r}')} = \frac{\delta^2 \ln \Xi[\psi]}{\delta \psi(\mathbf{r}) \delta \psi(\mathbf{r}')} = \rho(\mathbf{r})\rho(\mathbf{r}')h(\mathbf{r},\mathbf{r}') + \rho(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}'). \tag{43}$$

From which the total correlation function can be written as:

$$h^{(2)}(\mathbf{r},\mathbf{r}') = \frac{1}{\rho(\mathbf{r})\rho(\mathbf{r}')}\frac{\delta^2 \ln \Xi[\psi]}{\delta\psi(\mathbf{r})\delta\psi(\mathbf{r}')} - \frac{\delta(\mathbf{r}-\mathbf{r}')}{\rho(\mathbf{r}')}. \tag{44}$$

An analogous derivation for the direct correlation function yields

$$c^{(2)}(\mathbf{r},\mathbf{r}') = -\beta\frac{\delta^2 F^{ex}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')} \tag{45}$$

with the excess free energy functional $F^{ex}[\rho]$. Here $\Omega[\psi]$ and $F[\rho]$ are related by the generalized Legendre transformation

$$\Omega[\psi] - \int\psi(\mathbf{r}) \to \Omega[\psi]\int\psi(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} = F[\rho] \tag{46}$$

from which the Ornstein-Zernike equation (39) emerges.

For a general liquid, including the non-spherical molecules of interest in this thesis, the angular dependence of the total correlation function is considered by the Euler angles:[158]

$$h^{(2)}(\mathbf{r},\boldsymbol{\Omega},\mathbf{r}',\boldsymbol{\Omega}') = c^{(2)}(\mathbf{r},\boldsymbol{\Omega},\mathbf{r}',\boldsymbol{\Omega}') + \int c^{(2)}(\mathbf{r}',\mathbf{r}'',\boldsymbol{\Omega}',\boldsymbol{\Omega}'')\rho^{(1)}(\mathbf{r}'')h^{(2)}(\mathbf{r}',\mathbf{r}'',\boldsymbol{\Omega}',\boldsymbol{\Omega}'')d\boldsymbol{\Omega}d\mathbf{r} \tag{47}$$

Which then is called the molecular Ornstein-Zernike equation[161][162](MOZ). Here however both $h^{(2)}$ and $c^{(2)}$ are unknown requiring a further relation called the closure.

To overcome the issue of the high dimensionality of the MOZ Chandler and Andersen [163][164]developed a formalism based on spherical “interaction sites” to model an arbitrary, but rigid, solvent. For example, the direct correlation function can be approximated as a sum over pairwise interactions

$$c_{\alpha\gamma}(r) \approx \sum_{\alpha}\sum_{\gamma} c_{\alpha\gamma}(|\mathbf{r}_\alpha - \mathbf{r}_\gamma|). \tag{48}$$

The structure of the solvent molecule is retained by the intramolecular correlation function

$$\omega_{\alpha\gamma}(r) = \frac{\delta(r - l_{\alpha\gamma})}{4\pi l_{\alpha\gamma}^2} \tag{49}$$

with the intramolecular distance $l_{\alpha\gamma}$ between interaction sites α and γ. In this formulation, however the solvent molecule remains rigid and every small change in geometry results in a new iterative cycle of solving the MOZ.

The resulting RISM equation for the total correlation function can be written in matrix from as

$$\mathbf{h} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\rho}\boldsymbol{\omega} * \mathbf{c} * \mathbf{h} \,. \tag{50}$$

where $*$ denotes a convolution. It can be calculated separately for solvent-solvent interactions (*vv*), solute-solvent interactions (*uv*) or solute solute interactions (*uu*). However, both the MOZ and the RISM equation need another equation to be solved which is called the closure equation and brings physical properties into the description of the solvent. For a site model of a solvent the closure can be written as

$$h_{\alpha\gamma}(r) = \exp\left(-\beta u_{\alpha\gamma}(r) + h_{\alpha\gamma}(r) - c_{\alpha\gamma}(r) + B_{\alpha\gamma}(r)\right) - 1 \tag{51}$$

with the interaction potential $u_{\alpha\gamma}$, $\beta = \frac{1}{k_B T}$ and the so-called bridge function $B_{\alpha\gamma}$ which is an unknown function in RISM. To still be able to calculate the total correlation function within RISM different approximations can be used with one of the most common being the hypernetted-chain approximation[165] which in RISM sets the bridge function to zero. It is then possible to calculate the total correlation function or pair distribution function $g(r) = h(r) + 1$ at a given thermodynamic condition of pressure or temperature. While the temperature is directly encoded in the closure equation pressure can be varied by a change in particle density which corresponds to the solvent density at the given pressure. With this there is no formal difference between a high-pressure and an ambient-condition calculation as no additional terms accounting for pressure need to be included.

For the accurate description of the solvent it is then necessary to use accurate interaction potentials $u_{\alpha\gamma}$ with a combination of Lennard-Jones and electrostatic Coulomb potentials being among the most common. Parameters can then be taken from various force fields as done in this thesis.

The development of the 3D-RISM goes back to the 1990s and the developments of Beglov and Roux[166], and Kovalenko and Hirata[167] and considers the presence of a solute at infinite dilution. The resulting 3D RISM equation is

$$\rho_\gamma h_\gamma(\mathbf{r}) = \sum c_{\gamma'} * \chi_{\gamma\gamma'}(\mathbf{r}) \tag{52}$$

with $\chi_{\gamma\gamma'}(\mathbf{r})$ as the solvent susceptibility function. It can be precomputed from 1D-RISM via

$$\boldsymbol{\chi} = \boldsymbol{\rho}^v \boldsymbol{\omega}^v + \boldsymbol{\rho}^v \mathbf{h}^{vv} \boldsymbol{\rho}^v \,. \tag{53}$$

However, it can also be derived from simulations of the pure solvent.[168]

It allows for the introduction of specific thermodynamic conditions into 3D-RISM as it mimics the behavior of the solvent at these conditions.

A clear advantage of RISM is the possibility of calculating the excess chemical potential from equation (37) analytically as:[169]

$$\Delta\mu^{\text{ex}} = -\frac{\rho}{\beta}\sum_{\alpha,\gamma}\int\left(\frac{1}{2}h_{\alpha\gamma}(\mathbf{r})^2 - c_{\alpha\gamma}(\mathbf{r}) + \frac{1}{2}h_{\alpha\gamma}(\mathbf{r})c_{\alpha\gamma}(\mathbf{r})\right)d\mathbf{r} \tag{54}$$

within the HNC approximation. The excess chemical potential corresponds to the excess Helmholtz free energy in the limit of infinite dilution Due to numerical instabilities of the HNC-closure in 3D-RISM the so-called PSE-*n* closures were developed as a partial series expansion of the 3D HNC closure as[170]

$$h_\gamma(\mathbf{r}) = \begin{cases} \sum_{i=0}^{k}\left(-\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})\right)^i & \text{for } -\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) > 0 \\ \exp\left(-\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})\right) - 1 & \text{for } -\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) \le 0 \end{cases} \tag{55}$$

Here the PSE-1 closure is often referred to as the Kovalenko-Hirata (KH) closure.[171] The excess chemical potential using the PSE-closures is given by

$$\Delta\mu^{\text{ex}} = -\frac{\rho}{\beta}\sum_{\alpha,\gamma}\int\frac{\Theta h_{\alpha\gamma}(\mathbf{r})\left(h_{\alpha\gamma}(\mathbf{r}) - \beta u_{\alpha\gamma}(\mathbf{r}) - c_{\alpha\gamma}(\mathbf{r})\right)^{k+1}}{(k+1)!}d\mathbf{r}\,. \tag{56}$$

The interaction potential u_γ is similar to the nonbonded term in equation (23):

$$u_\gamma(\mathbf{r}) = \sum_\alpha \frac{q_\alpha q_\gamma}{4\pi\varepsilon_0|\mathbf{r}-\mathbf{R}_\alpha|} + 4\varepsilon_{\alpha\gamma}\left[\left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{R}_\alpha|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{R}_\alpha|}\right)^{6}\right] \tag{57}$$

containing a term for the Coulomb potential and the Lennard-Jones potential.

Analogous to MD simulations to describe the solute-solvent interaction Lennard-Jones parameters $\varepsilon_{\alpha\gamma}$ and $\sigma_{\alpha\gamma}$ get mixed, most of the time using arithmetic mixing rules called Lorentz-Berthelot mixing rules.[172] For a quick calculation of the long range Coulomb potential with minimal errors due to finite box size effects a renormalization approach can be used.[173] Here the electrostatic Coulomb potential gets split into a short-range term

$$u_\gamma^{\text{elec,short}}(\mathbf{r}) = \sum_\alpha \frac{q_\alpha q_\gamma}{4\pi\varepsilon_0|\mathbf{r}-\mathbf{R}_\alpha|}\operatorname{erfc}\left(\kappa|\mathbf{r}-\mathbf{R}_\alpha|\right) \tag{58}$$

and a long-range term

$$u_{\gamma}^{\text{elec,long}}(\mathbf{r}) = \sum_{\alpha} \frac{q_\alpha q_\gamma}{4\pi\varepsilon_0 \left|\mathbf{r}-\mathbf{R}_\alpha\right|} \operatorname{erf}\left(\kappa\left|\mathbf{r}-\mathbf{R}_\alpha\right|\right). \tag{59}$$

Here the short-range term can be directly calculated in the real space while the long range term is efficiently calculated using a Fourier transform with a proper smearing factor κ. This avoids the expensive calculation of real-space electrostatic terms on the 3D grid.

2.4.4 The combination of quantum mechanics and RISM

With the pair interaction potentials appearing in equation (57) there are multiple ways to calculate the individual terms. Here using terms from classical force fields for both the electrostatic and Coulomb potential is efficient but combining QM calculations with the RISM approached showed a large improvement in accuracy. Current implementations of 3D-RISM together with QM calculations are able to reach so-called chemical accuracy (1 kcal/mol) in the prediction of solvation free energies in various solvents.[174][175][176] Similar accuracies can be reached for the prediction of reaction equilibria in the form of pK_a values.[177] Successful applications to spectroscopic parameters such as NMR[178][179] or EPR parameters[9] have also been made.

This broad range of applications is a result of multiple approaches that have been developed over the years that derive the partial charges within the electrostatic potential not from force fields but from quantum mechanical calculations. At first the RISM-SCF was developed by Ten-No et al. in the early 1990s.[180][181] In its first use cases RISM-SCF used the QM derived partial charges of the solute which were plugged into a 1D-RISM calculation. The resulting solvent distribution from 1D-RISM was then added to the Fock-Operator in the QM calculation representing a polarizing solvent. The idea was later extended to 3D-RISM calculations and Kohn-Sham DFT by Kovalenko and Hirata using first plane-wave basis sets[182] and later Gaussian basis sets.[183] Later implementations of RISM-SCF[184][185] then relied on the use of the wavefunction-based exact electrostatic potential of the solute instead of partial charges. Variations of 3D-RISM-SCF have been developed with 3D-RISM-SCF-SEDD[186] where the solute point charges are replaced by a linear combination of auxiliary basis sets thereby introducing the spatial electron density distribution (SEDD) and 3D-RISM-SCF-cSED where additional constraints of the spatial electron density (cSED) are introduced for more stable calculations.[187][188] New technical implementations of 3D-RISM-SCF are still developed highlighting the importance of this method to describe solvent phenomena. In 2020 Reimann and Kaupp [178] added an implementation to the Amsterdam Density Functional (ADF) program[189] that uses the exact electrostatic potential for 3D-RISM calculations and in 2024 the RISMiCal program package[190] was developed having fast GPU code for RISM calculations with a direct interface to common QM programs.

Apart from the 3D-RISM-SCF methodology the group of Kast developed the EC-RISM[33]. The scheme of an EC-RISM calculation is shown in Figure 5.

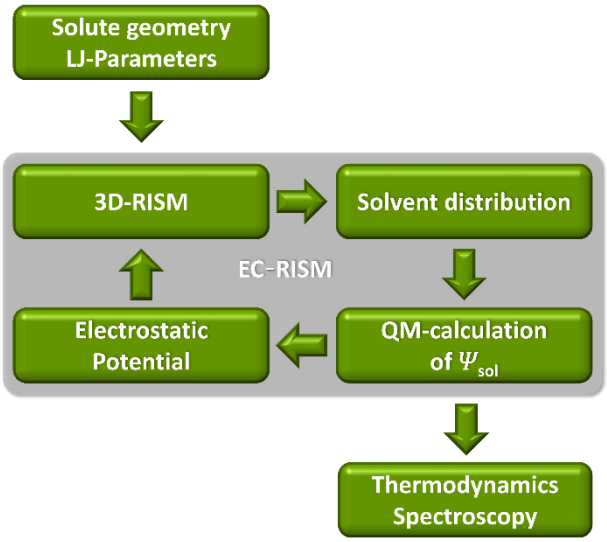

Figure 5: Schematic representation of an EC-RISM calculation. Based on an initial solute geometry with Lennard-Jones (LJ) parameters a cycle of a 3D-RISM and a QM-calculation is performed until convergence. From the solvated wavefunction thermodynamic or spectroscopic observables can be calculated.

In the same way as 3D-RISM-SCF EC-RISM takes into account the polarization of solvent by the solute and polarization of the solute by the solvent in a self-consistent way until a converged wave function is reached. By combining 3D-RISM with quantum mechanics thermodynamic properties such as the free energy of the solute in solution G^{sol} can be calculated as

$$G^{sol} = E^{sol} + \mu^{ex} \tag{60}$$

from the QM-derived solute energy E^{sol} using the solvated wave function ψ_{sol}

$$E^{sol} = \left\langle \psi_{sol} \middle| \hat{H}_{ne} + \hat{H}_{ee} + \hat{H}_{nn} \middle| \psi_{sol} \right\rangle \tag{61}$$

and the excess chemical potential μ^{ex} which can be calculated from equation (54). As a result of the calculation EC-RISM yields a wave function representing a solute in solution. From the wave function spectroscopic observables such as NMR or EPR parameters can be calculated.[179][9] The theory behind the calculation of NMR and EPR parameters from wave function based methods will be explained in the following chapter.

To further illustrate the EC-RISM cycle in Figure 5 necessary ingredients are Lennard-Jones parameters of the solute which for a broad range of small molecules can be taken from force fields such as the general amber force field (GAFF).[121] As another ingredient for an EC-RISM calculation the 3D-RISM

calculation needs a precomputed solvent site susceptibility function $\boldsymbol{\chi}$ derived from equation (53) within 1D-RISM.

The calculation then starts with QM calculation in vacuum from which the electrostatic potential can be derived. This is then used in a 3D-RISM calculation polarizing the solvent. The result of the 3D-RISM calculation is the solvent distribution g_γ of each site, which is then, represented by point charges, included in the next QM calculation to represent the polarizing solvent. The point charges can be calculated using

$$\rho_{\mathrm{q}}(\mathbf{r}) = \sum_{\gamma} q_\gamma \rho_\gamma g_\gamma(\mathbf{r}) \tag{62}$$

with the partial charges of the solvent q_γ which can be taken from solvent specific force fields. These point charges can then be used in the next QM calculation continuing the EC-RISM cycle. This is repeated until convergence is reached.

When using 3D-RISM and therefore also EC-RISM to calculate thermodynamic properties such as the solvation free energy large discrepancies originally showed up. It was found that these scale with the partial molar volume V_{m} of the solute indicating an overestimation of the pressure within 3D-RISM which lead to an empirical, so-called PMV correction, of the excess chemical potential[174]

$$\mu^{\mathrm{ex,corr}} = \mu^{\mathrm{ex}} + c_{V_{\mathrm{m}}} V_{\mathrm{m}} + c_q q\ . \tag{63}$$

Here it is noteworthy that the c_{Vm} parameter directly depends on the compressibility of the solvent for which experimental values can be taken. It can also be calculated within RISM.[191] The PMV correction also includes the c_{q} parameter necessary to modify the excess chemical potential of net-charged species. The physical reason for this parameter is the neglection of the Galvani potential for the determination of the proton hydration free energy and derived ion hydration free energies.[192] In addition however, it has to be noted that even experimental ion solvation free energies are based on a chosen absolute reference value which is, most commonly, the value for the absolute hydration free energy of a proton by Tissandier.[193] The c_{q} parameter is then also able to correct for potential mismatches in the absolute free energy scale based on the chosen reference value. A c_{q} is not necessary for apolar species like cyclohexane and calculation of distribution coefficients with apolar solvents for which it can be assumed that charged species are not present in these solvents.

Other correction schemes to account for the overestimation of pressure within 3D-RISM include the so-called universal correction[194] or the PC+ correction.[195]

Additional model-specific parameters modifying the calculated excess chemical potential or adding an additional offset can be applied depending on the model,[177] however these are fully empiric parameters.

To determine these parameters datasets like the Minnesota solvation database (MNSOL)[196] can be used. These contain experimental data for the solvation free energy which, from EC-RISM, can be expressed, using the so called Ben-Naim reference state[197] for infinite dilution as[174]

$$\Delta_{\text{solv}}G^0 = \Delta_{\text{reorg}}E + \Delta_{\text{solv}}E + \mu^{\text{ex,corr}} \tag{64}$$

with the reorganization energy $\Delta_{\text{reorg}}E$, and the solvation energy $\Delta_{\text{solv}}E$ which can both be calculated from the difference between the gas phase ensemble and the solution phase ensemble. For this equation it is assumed that both the gas phase ensemble and the solution phase ensemble consist of only one structure and rotational and vibrational entropy contributions are disregarded. Here $\Delta_{\text{reorg}}E$ is the difference in the intramolecular energy between solution and vacuum structure computed in the gas phase. $\Delta_{\text{solv}}E$ represents the solute's polarization energy change upon transfer from gas phase into solution.

Under the assumption that both ensembles are the same, which would mean that no structural or polarization change takes place in solution compared to gas phase, which is a force field type approach, the solvation free energy can also be estimated by the (corrected) excess chemical potential

$$\Delta_{\text{solv}}G^0 \approx \mu^{\text{ex,corr}} . \tag{65}$$

As the partial molar volume correction is based on an overestimation of the pressure within the RISM framework pressure-dependent calculations amplify this error. Within RISM high pressure is introduced by a change of the solvent susceptibility, which introduces potential changes of the solute geometry, density and electronic permeability under high-pressure conditions. It is then possible to perform high-pressure which just requires another, pressure-dependent correction which can be written as

$$\mu^{\text{ex,corr,HP}} = \mu^{\text{ex,corr}} + c_{\text{HP}}(p - 1bar)V_{\text{m}}(p) \tag{66}$$

including the additional parameter c_{HP}. Here the use of the Ben Naim reference state to define the solvation free energy is especially useful as the thermodynamic standard of 1 M implies a pressure-dependent volume change which is not present for an infinite dilution standard state.

As experimental data for pressure-dependent solvation free energies is very limited reference data needs to be computed by methods such as thermodynamic integration (TI).[198] The pressure-dependent solvation free energy can then be used to reference data as:[8]

$$\Delta_{\text{solv}}G^0{}_{\text{ref}}(p) = \Delta_{\text{solv}}G^0{}_{\text{TI}}(p) + \Delta_{\text{solv}}G^0{}_{\text{EC-RISM}}(1\text{ bar}) - \Delta_{\text{solv}}G^0{}_{\text{TI}}(1\text{ bar}) \tag{67}$$

This also accounts for possible differences between EC-RISM and TI at ambient conditions.

As one of the most recent developments within EC-RISM it can be used for geometry optimizations. For that a direct implementation of EC-RISM in the quantum chemistry package ORCA[199] was realized by P. Kibies and M.A. Garcia-Rates. To perform geometry optimizations the calculation of a gradient is necessary which for the free energy can be written as

$$\frac{\partial G}{\partial \mathbf{R}_\alpha} = \frac{\partial E_{\text{sol}}}{\partial \mathbf{R}_\alpha} + \frac{\partial \mu^{\text{ex}}}{\partial \mathbf{R}_\alpha} \tag{68}$$

with the gradient of the intramolecular energy which can be accessed directly from the quantum chemistry program and the gradient of the access chemical potential which can be written as[200]

$$\frac{\partial \mu^{\text{ex}}}{\partial \mathbf{R}_\alpha} = \sum_\gamma \rho_\gamma \int g_\gamma(\mathbf{r}) \left(\frac{\partial u_\gamma^{\text{LJ}}}{\partial \mathbf{R}_\alpha} + \frac{q_\gamma \partial \phi(\mathbf{r}, \mathbf{R}_\alpha)}{\mathbf{R}_\alpha} \right) d\mathbf{r} \,. \tag{69}$$

2.5 Introduction to magnetic resonance spectroscopy

Magnetic resonance spectroscopy is one of the most used tools in analytical chemistry to determine chemical structures of small molecules and large macromolecules such as proteins. It can give very accurate insights into molecules on an atomistic level. In this chapter the basics to understand experimental NMR spectroscopy as well as electron paramagnetic resonance (EPR) spectroscopy will be explained.

The basis of both spectroscopy methods was discovered by Zeeman in 1896 and first observed in the 1940s by Bloch[201] and Purcell[202] (NMR), and Zavoisky[203] (EPR). An object with a spin experiences an energetic splitting (Zeeman-splitting) within an external magnetic field B_0. For a spin with the spin magnetic moment m_I and gyromagnetic ratio γ_I the energy is given by

$$E_{m_\text{I}} = -m_\text{I} \gamma_\text{I} \hbar B_0 \tag{70}$$

with the reduced Planck constant $\hbar$. For electrons the magnetic moment can be determined by:

$$\mathbf{m}_\text{I} = \gamma_\text{s} \mathbf{s}_\text{I} = -g_e \frac{\mu_\text{B}}{\hbar} \mathbf{s}_\text{I} = -g_\text{e} \frac{e}{2m_\text{e}} \mathbf{s}_\text{I} \tag{71}$$

where $\mathbf{s}_\text{I}$ is the spin vector μ_B is the Bohr magneton, e is the elementary charge, m_e is the mass of the electron, g_e is the Landé-factor of the electron. For nuclear spins *c* the analogous equation is:

$$\mathbf{m}_\text{I} = \gamma_\text{c} \mathbf{I}_\text{c} = g_c \frac{\mu_\text{N}}{\hbar} \mathbf{I}_\text{c} = g_\text{c} \frac{e}{2m_p} \mathbf{I}_\text{c} \tag{72}$$

with the mass of the proton m_p .Here the sign changes due to the opposite charge in the nucleus compared to electrons. The gyromagnetic ratio of the nucleus is highly dependent on the specific nucleus in question which leads to large changes in the energetic splitting of different nuclei. The nuclear spin vector is given by:

$$|\mathbf{I}_c| = \sqrt{I(I+1)\hbar^2} \tag{73}$$

with the nuclear spin number *I* which depends on the number of protons and neutrons in the nucleus. For an even number of protons and neutrons *I* is zero and the nucleus can not be measured in NMR spectroscopy. For nuclei with an odd mass number *I* is a half number similar to the electron which has a spin quantum number of ½. For nuclei with an even mass number and odd number of protons and neutrons *I* is an integer. The nuclear spin number has a direct impact on the Zeeman splitting as $\mathbf{m}_I$ can reach in integer steps from -*I* to +*I* leading to a multiplicity *M* of:

$$M = 2I + 1 \tag{74}$$

A spectroscopic signal is then dependent on the ratio of populations between to states which can derived from Boltzmann statistics as

$$\frac{N_2}{N_1} = \exp(-\beta \Delta E) \tag{75}$$

with $\beta = \frac{1}{k_B T}$. The energetic differences between the two states of an electron spin or a nuclear spin of I = ½ can then be written as

$$\Delta E_{m_I} = \gamma_I \hbar B_0 = g_I \mu_I B_0 \tag{76}$$

which is much larger for EPR spectroscopy compared to NMR spectroscopy due to the higher gyromagnetic ratio of the electron.

As the main idea of the spectroscopy resonance can be achieved with an electromagnetic wave exactly matching this energy difference. The frequency ν of this wave is then

$$\nu = \frac{\gamma_I B_0}{2\pi} \tag{77}$$

which is also called the Larmor frequency.

The essence and the reason for the success of NMR and EPR spectroscopy is that not only an external magnetic field can lead to Zeeman splitting. With other nuclei and electrons also being spins they

modulate the so called effective magnetic field for their surrounding atoms or electrons leading to different signals based on the specific chemical structure of an atom or unpaired electron within a molecule. The resulting effective magnetic field B_{eff} can be written as

$$B_{\text{eff}} = (1-\sigma)B_0 \tag{78}$$

with the shielding constant σ which is used in the context of NMR or as

$$B_{\text{eff}} = (g / g_{\text{e}})B_0 \tag{79}$$

with the *g*-value which is used in the context of EPR. This modifies the resonance condition (77) to e.g.

$$\nu = \frac{\gamma_{\text{I}}}{2\pi} B_0(1-\sigma)\,. \tag{80}$$

As a result of different spectrometers in use for NMR spectroscopy an observable independent of the magnetic strengths was desirable. This led to the definition of the chemical shift δ which is defined by the relative change in resonance frequency or change in shielding constant compared to a reference system according to

$$\delta = \frac{\nu - \nu_0}{\nu_0} \cdot 10^6 = \frac{(1-\sigma)B_0 - (1-\sigma_0)B_0}{(1-\sigma_0)B_0} \cdot 10^6 \approx (\sigma_0 - \sigma) \cdot 10^6\,. \tag{81}$$

The chemical shift is then given in ppm. For the most common nuclei in NMR spectroscopy ^{1}H and ^{13}C tetramethylsilane (TMS) is used as a reference in nonpolar solvents. For measurements in aqueous solutions the methyl groups of sodium trimethylsilylpropanesulfonate (DSS) are often used as a reference.[204] For ^{15}N liquid ammonia in an external capillary can be used as a reference. An alternative unified referencing scale uses the so called Ξ-factors which relates every resonance frequency back to the ^{1}H of TMS or DSS by[205]

$$\Xi = \frac{\nu_{\text{x}}^{\text{obs}}}{\nu_{\text{TMS/DSS}}}\,. \tag{82}$$

This method of referencing can be used for a wide variety of nuclei for which data for Ξ-factors is available.[206][207] With this method of referencing no internal standard is necessary and only the ^{1}H resonance frequency of TMS or DSS needs to be determined at the used spectrometer.

This method is also applicable for high pressure calculations, but it assumes a pressure-independent chemical shift of both the ^{1}H reference (TMS or DSS) and the internal reference nucleus of interest to be comparable to the direct reference. A deeper analysis of this for the case of ^{13}C using DSS will be

performed in this thesis. A use of pressure-dependent Ξ-factors could also overcome potential pressure-dependent chemical shift changes of the reference nucleus, but has not been done before.

In addition to the chemical shift in NMR spectroscopy and the *g*-value in EPR spectroscopy which modify the resonance frequency depending on the chemical environment an additional splitting of signals can be observed. This is due to the interactions of nuclear spins with the nucleus (NMR) or electron (EPR) in question. In EPR this is called hyperfine coupling and modifies equation (76) to[208]

$$\Delta E_{m_I} = g_I \mu_I B_0 \pm \frac{m_I}{2} a \tag{83}$$

with the hyperfine coupling constant a. The analogous case in NMR spectroscopy is called *J*-coupling. Depending on the nuclear spin number the number of lines that are observable in the spectrum because of hyperfine- or *J*-coupling again follow equation (74) and their intensities follow the famous Pascal's triangle. Hyperfine- or *J*-coupling can be attributed to a so-called dipolar coupling which refers to the direct interaction of two magnetic dipols through space and a trough-bond interaction which is also called the Fermi contact interaction.[209]

From a practical point of view one difference between NMR and EPR spectroscopy is noteworthy. As NMR spectroscopy is usually performed in a liquid solution at or around room temperature specific orientations of nuclear spins average out over the measurement time of milliseconds to seconds leading to an isotropic measurement. This is in contrast to EPR spectroscopy which is usually performed for frozen samples at very low temperatures to improve signal intensities following equation (75). Therefore, in an EPR spectrum anisotropy of the g-value and hyperfine coupling constant due to the dipolar coupling can be observed. Even though the anisotropy of the dipolar coupling is usually not observed in NMR spectroscopy it is still responsible for the Nuclear Overhauser Effect (NOE) and measurements of exploiting the NOE (e.g. NOESY)[210] are frequently used to determine intra- or intermolecular distances.

2.6 Theoretical calculation of magnetic resonance parameters

The goal of this chapter is to show ways to calculate the experimental observables described in the previous chapter using quantum chemical methods of which some have been described in chapter 2.1. The main way to calculate magnetic resonance parameters is called second order perturbation theory and was developed by Ramsey in the 1950s.[211][212][213] To explain the basics of this theory first the Hamiltonian that describes the spin system needs to be derived. The system is modelled by two states *a* and *b* between which a NMR- or EPR-measurable transition takes place. The Hamiltonian that accurately represents this system should be able to account for the transition between the two states. These states can be written as[214]

$$\Phi^a = \left|\Psi, m_s^a, m_1^a \ldots m_N^a\right\rangle \text{ and } \Phi^b = \left|\Psi, m_s^b, m_1^b \ldots m_N^b\right\rangle \tag{84}$$

with the spatial electronic wave function Ψ an electronic spin state described by its quantum number m_s and nuclear spin state defined by $m_1 \ldots m_N$. As both states have the same spatial wave function the energy difference corresponding to the resonance condition (e.g. equation (76)) is

$$\Delta E^{ab} = E\left(\Phi^b\right) - E\left(\Phi^a\right) = E\left(m_s^b, m_1^b \ldots m_N^b\right) - E\left(m_s^a, m_1^a \ldots m_N^a\right) \tag{85}$$

and only depends on the spin numbers of states a and b. To extract these energy values a corresponding Hamiltonian is necessary which is called the effective spin Hamiltonian $\hat{H}_s$. It effectively takes into account the contributions of other electrons surrounding the electron (EPR) or nucleus (NMR) in question. The energy of a given state can then be calculated solely based on the magnetic moments via

$$\begin{aligned}\left\langle \Phi^a \middle| \hat{H}_s \middle| \Phi^a \right\rangle &= \left\langle m_s^a, m_1^a \ldots m_N^a \middle| \hat{H}_s \middle| m_s^a, m_1^a \ldots m_N^a \right\rangle \\ &= E\left(m_s^a, m_1^a \ldots m_N^a\right).\end{aligned} \tag{86}$$

The exact formulation of the Hamiltonian is based on conventions that have evolved over time and might differ between communities. However, they should take into account the basic principle of magnetic resonance spectroscopy namely the interaction of a spin with an external magnetic field. A common spin Hamiltonian for NMR spectroscopy is[214]

$$\hat{H}_s(\text{NMR}) = -\sum_{C=1}^{N} \hbar\gamma_C \mathbf{B}\cdot\left(\mathbf{1} - \boldsymbol{\sigma}_c\right) + \frac{\hbar^2}{2}\sum_{C=1}^{N}\sum_{\substack{D=1 \\ D\neq C}}^{N} \gamma_C\gamma_D \mathbf{I}_C \cdot \left(\mathbf{D}_{CD} + \mathbf{J}_{CD}\right)\cdot \mathbf{I}_D \tag{87}$$

where $\mathbf{1}$ is the unit tensor, $\boldsymbol{\sigma}_c$ is the nuclear magnetic shielding tensor, $\mathbf{D}_{CD}$ is the dipolar coupling tensor and $\mathbf{J}_{CD}$ the indirect nuclear spin-spin coupling tensor. For EPR spectroscopy an often used spin Hamiltonian is

$$\hat{H}_s(\text{EPR}) = \mu_B \mathbf{S}\cdot\mathbf{g}\cdot\mathbf{B} + \sum_{C=1}^{N} \mathbf{S}\cdot\mathbf{A}_C\cdot\mathbf{I}_C \tag{88}$$

where the first term describes the Zeeman splitting analogous to equation (70) and the second term describes the hyperfine coupling introduced in equation (83).

Now that the spin Hamiltonians are defined second order perturbation theory describes how to extract eigenvalues of the specific tensors within the Hamiltonian by (mixed) second order derivatives. In general these second order derivatives can be written in the sum over states formalism as e.g.:[214]

$$\left(\frac{\partial^2 E}{\partial\kappa\partial\lambda}\right)=E_{11}=\langle\psi_0|\hat{H}_{11}|\psi_0\rangle+\sum_{p>0}\frac{\langle\phi_0|\hat{H}_{10}-E_{10}|\phi_p\rangle\langle\phi_p|\hat{H}_{01}-E_{01}|\phi_0\rangle}{\varepsilon_0-\varepsilon_p} \tag{89}$$

Here it is written as sum over excited states p normalized by the difference in energy ε compared to ground state. Indices are used to describe the perturbation order with respect to the given parameter. For example E_{11} as a second-order property represents the first derivate with respect to the first parameter
(κ) and the second parameter (λ)

To now calculate experimental observables the following second derivatives have to be calculated. The nuclear shielding tensor can be calculated as the derivative with respect to the magnetic moment and magnetic field:

$$\boldsymbol{\sigma}=\left(\frac{\partial^2 E}{\partial\boldsymbol{\mu}\partial\mathbf{B}}\right) \tag{90}$$

It can be related to a measured chemical shift using equation (81). As the chemical shift in liquid solution is an isotropic value the isotropic shielding constant has to be calculated using the Trace of the shielding tensor

$$\sigma=\frac{1}{3}\left(\sigma_{xx}+\sigma_{yy}+\sigma_{zz}\right)=\frac{1}{3}\mathrm{Tr}\left(\boldsymbol{\sigma}\right). \tag{91}$$

The magnetic susceptibility tensor can be calculated as

$$\boldsymbol{\chi}=\left(\frac{\partial^2 E}{\partial\mathbf{B}_a\partial\mathbf{B}_b}\right) \tag{92}$$

while the nuclear spin coupling tensor is defined as

$$J=\left(\frac{\partial^2 E}{\partial\boldsymbol{\mu}_a\partial\boldsymbol{\mu}_b}\right). \tag{93}$$

The g-tensor can be calculated in analogous ways using

$$\mathbf{g}=\left(\frac{\partial^2 E}{\partial\mathbf{B}\partial\mathbf{s}}\right) \tag{94}$$

with the change that the magnetic moment is that of an electron and not of a nucleus. Here individual tensor components might be of interest as they are also experimentally measurable, but isotropic values for all of these tensors can be calculated using the trace of the tensor.

One exception is the hyperfine coupling constant as it can be calculated as a first-order parameter based on

$$\mathbf{A} = \langle \psi_0 | \hat{H}_{fc} + \hat{H}_{sd} | \psi_0 \rangle . \tag{95}$$

Taking the Fermi contact interaction and spin-dipole interaction into account. Which both only depend on the magnetic moment and therefore only need to be treated using first order perturbation. From a practical point of view in perturbation theory these derivatives are calculated around the unperturbed wave function at zero-field as shown in equation (89). This is justified as the induced current density in molecules scales approximately linear up to the experimentally used magnetic field.[214] Still, a problem arises in practical calculations using finite basis sets calculated properties depend on the chosen gauge of the basis as the chose gauge for example directly influences $\hat{H}_{sd}$. With increasing basis sets the gauge dependence decreases up to an infinite basis set where calculations are once again exact. However, with finite basis sets used most often in real calculations and also solely used in this thesis methods to overcome the gauge dependence were necessary and have been developed. These methods include the individual gauge for localized orbitals (IGLO),[215][216] the continuous set of gauge transformations (CSGT),[217] the individual gauges for atoms in molecules (IGAIM)[218] and most prominently the gauge invariant atomic orbitals (GIAO)[219][220] which was also used within this thesis. These orbitals have the form[214]

$$\chi_\nu = \phi_\nu \exp\left(-\frac{ie}{\hbar c} \mathbf{A}_i \cdot \mathbf{r} \right) \tag{96}$$

which modifies the basis function ϕ_ν with it's coefficient c by the vector potential $\mathbf{A}_i$. For the calculation of magnetic properties a vector field describing a magnetic field is given by:

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{97}$$

So that the curl of the vector potential is the magnetic field. As shown before this potential is differentiated during the calculation of magnetic properties which allows the addition of a constant which represents the gauge origin of the potential.[221] This makes the wave function built on the orbitals shown in equation (95) explicitly dependent on the chosen gauge origin, but magnetic properties derived from it are gauge invariant allowing for calculations using approximate basis sets.

Matching theoretical NMR calculations with experimental results also requires the use of a consistent referencing method. Here the direct reference using the same compounds as in the experiment is the straightforward method ensuring the highest comparability between experiment and calculations.

However, due to the potential difference in the chemical composition of the reference and target compound, different methods have been developed to benefit from error cancellation within NMR calculations.

A multi-standard approach makes use of one or many different referencing compounds and their experimental chemical shifts to calculate the, then smaller, chemical shift differences with respect to the chosen new reference. For example, Sarotti and Pellegrinet[222] used methanol as a reference for all sp^3 carbons and benzene for all sp and sp^2 carbons. This drastically improved calculated ^{13}C chemical shifts compared to direct reference to TMS making use of an improved error cancellation.

This approach can be taken even further by not calculating chemical shifts directly but instead fitting the calculated nuclear magnetic shielding constants to experimental chemical shifts.[30] The predicted chemical shifts are then calculated as

$$\delta = \frac{b - \sigma}{\mathrm{m}} \tag{98}$$

where *b* is the fitted intercept of the linear regression corresponding to a reference shielding constant and *m* is the negative slope.

This approach, however, requires the fit to large experimental datasets. These large datasets do not exist for the high-pressure NMR spectroscopy requiring a direct reference approach. The only way to use an indirect approach in the calculation and match the calculation to an indirect reference approach is to only indirectly reference the calculated pressure-dependent chemical shift change. This was previously done by L. Eberlein[223] to calculate reference shielding constants for ^{13}C and ^{15}N from EC-RISM calculations. In this way the pressure-dependent chemical shift change from ^{1}H DSS is transferred to different nuclei. Compared to the directly calculated chemical shift change the indirect method yielded a lower chemical shift change, due to the relative pressure-insensitivity of ^{1}H DSS which was also found experimentally.[7]

Analyzing how these different reference method impact both calculated and experimental chemical shifts, especially for ^{13}C under high-pressure conditions, will be part of this thesis. The goal is to give recommendations for maximum consistency between experiments and calculations.

3. Methods

3.1 Pressure-dependent NMR measurements

With the use to benchmark and improve the accuracy of quantum chemical calculations in mind an experimental benchmark dataset consisting of 14 small molecules, shown in Figure 6, was created. The molecules were chosen, as they are small, soluble in water, ^{15}N-labelled and biologically relevant. A special focus was set on a group of small conformationally inflexible amides, which contains urea, acetamide, N-methylacetamide and benzamide. This so-called amide subset will be used for theoretical investigations, extending previous investigations of two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe.[224]

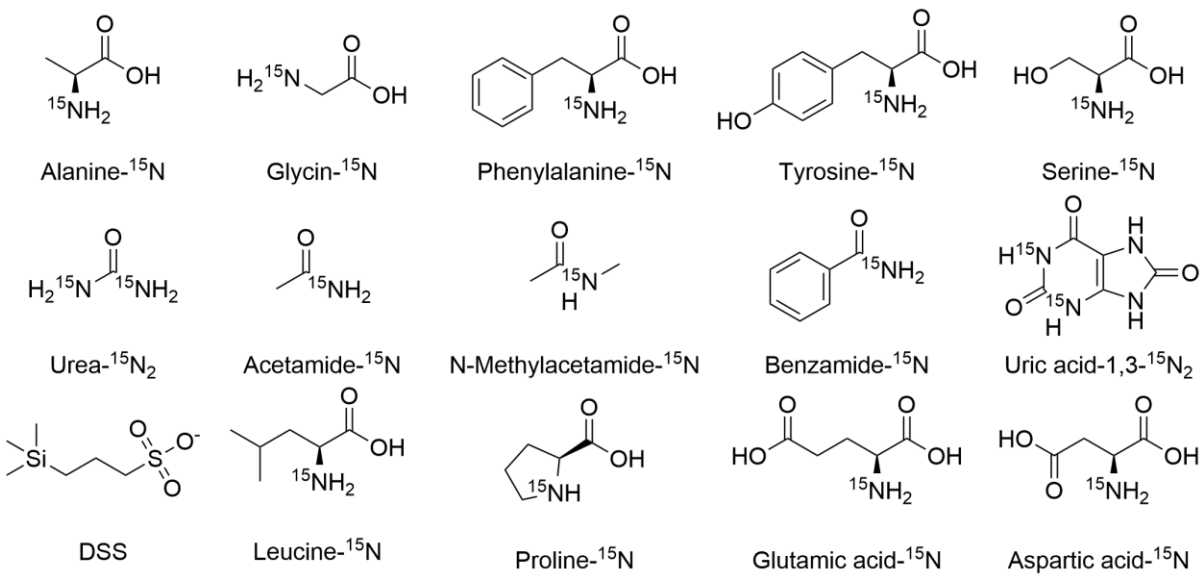

Figure 6: Molecules in the high-pressure dataset. DSS was used as an internal reference for ^{1}H and ^{13}C. For ^{15}N an indirect reference with Ξ= 0.101329118 (liquid NH_3 in a capillary immersed in a DSS solution) according to IUPAC recommendations was used.[206] The amide subset consists of urea, acetamide, N-methylacetamide and benzamide.

In total the selected molecules contain 127 NMR distinguishable nuclei with 58 ^{1}H, 52 ^{13}C, and 17 ^{15}N signals. Low solubility of uric acid and phenylalanine in water however made especially the ^{13}C and ^{15}N measurements for these compounds difficult, yielding unreliable data in these cases. The dataset contains fairly rigid molecules with low conformational flexibility along with flexible compounds, where a variety of tautomers and conformers are expected to be present in solution. This allows to investigate the intrinsic error of pressure-dependent NMR calculations as well as how tautomeric and conformational changes get influenced by high pressure. Here the ^{15}N-signals are expected to be most

sensitive to especially tautomeric changes involving the NH protons and were therefore chosen as a main feature of the benchmark dataset. The samples were prepared with 0.2 M ^{15}N-labelled molecule, 0.1 % (m/m) sodium salt of trimethylsilylpropanesulfonate (DSS), 0.1 M Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl) buffer, in 90% H_2O and 10% D_2O in a 5 mm outer diameter zirconia tube (540 µL volume) from Daedalus Innovations. The tube was filled with 450 µL of sample solution and 40 µL of mineral oil were added on top, while the remaining 50 µL were filled with H_2O. The hydrostatic pressure was applied to the sample directly in the magnet using the Xtreme 60 Syringe Pump from Daedalus Innovations. Pressure was transmitted by water while the mineral oil was used to prevent diffusion of the sample. The hydrostatic pressure was increased in steps of 500 bar up to a maximum of 3000 bar. After each pressure jump a time of 15 minutes passed until the start of the measurement to allow the sample to reach equilibrium conditions.

NMR experiments were performed with a Bruker Avance III HD NanoBay 400 MHz spectrometer equipped with a 5 mm room temperature broadband probe (BBFO Smart). Initial experiments for comparison were also performed using a Agilent DD2 500 MHz spectrometer. All experiments were carried out at 298.15 K. The ^{1}H spectra were recorded with a 30° flip angle and two steady-state scans followed by 16 scans. The acquisition time was 4.1 s with a relaxation delay of 5 s using 65536 data points. Proton spectra acquired with presaturation of the HDO signal used the zgpr pulse sequence, a 90° pulse, 16 steady-state scans, and 16 scans. The acquisition time was 5.2 s with a relaxation delay of 2 s using 66560 data points. The ^{13}C experiments were performed using power-gated proton decoupling with a 30° flip angle, 4 steady-state scans, 1024 scans, and an acquisition time of 1.36 s with a relaxation delay of 2 s using 65536 data points. All ^{15}N experiments were measured using inverse-gated proton decoupling, a 30° flip angle, and 64 or 128 scans depending on the concentration of the sample used. The FIDs were recorded with an acquisition time of 2.52 s and a 10 s relaxation delay using 65536 data points. The 90° pulse length was 9.4 µs for ^{1}H, 10.1 µs for ^{13}C, and 21.0 µs for ^{15}N. Data were processed, analyzed and plotted using MestReNova.[225]

^{1}H and ^{13}C resonances were directly referenced to the methyl resonance of DSS at 0 ppm. ^{15}N spectra were indirectly referenced to DSS using a Ξ-value of 0.101329118 corresponding to the use of liquid ammonia as a reference compound.[206]

3.2 FFMD simulations

A previous review[135] recommended the use of AIMD simulations to generate structural ensembles for improved NMR parameter predictions. As AIMD simulations however are too computationally demanding, FFMD simulations will be investigated in terms of their structural ensembles. For specific examples, namely DSS and NMA, in *cis*- and *trans* conformation, AIMD simulations were performed by B. Sharma and used as a benchmark for FFMD simulations. In addition, previous AIMD simulations of TMAO were reused as a further benchmark.[168]

All FFMD simulations were carried out using the Gromacs 2016.3 software package.[226] For ambient condition simulations solutes were placed in a box of 152 water molecules parameterized using the TIP4P-2005 force field.[227] For DSS, force field parameters were taken from the LigParGen webserver[228] using OPLS/AA[229] parameters with CM1A[230] charges while NMA as a model for the protein backbone was parametrized using the ff14SB[231] force field. All other solutes were parameterized using the GAFF 1.4 force field[121] on the ACPYPE web server.[232] Parameters for unlike atom pairs were described using Lorentz-Berthelot mixing rules for NMA and TMAO and geometric mixing for DSS. The van der Waals interactions were described with a 6-12 Lennard-Jones potential, truncated at 7.7 Å, and the electrostatic interactions were calculated with the smooth particle-mesh Ewald summation[233] with a real-space cutoff of 7.7 Å and a lattice spacing of 1 Å. A time step of 1.5 fs was used during *NpT* simulations, while all bonds including hydrogen were constrained using LINCS.[234] A temperature of 298.15 K was maintained by the stochastic velocity-rescaling thermostat,[126] and for ambient conditions a pressure of 1 bar was controlled by the isotropic Berendsen barostat[125] using a compressibility of $4.5 \cdot 10^{-5}$ bar^{-1}.

The production workflow for all FFMD simulations then contained 7.5 ns of *NpT* simulations where the last 6 ns were used to determine the density for further *NVT* simulations using FFMD. From the FFMD the snapshot closest to the average density was used as a starting point for further *NVT* simulations.

A density of 1016.55 g cm^{-3} was determined for DSS corresponding to a cubic supercell of 16.8601 Å while a density of 998.074 g cm^{-3} was found for NMA corresponding to a cubic supercell of 16.7246 Å. These densities were then also used for the AIMD simulations performed by B. Sharma.

For TMAO a density of 1000.8 g cm^{-3} corresponding to a cubic supercell of 16.7134 Å was obtained. It is noteworthy that the FFMD simulation of TMAO was performed in line with the other FFMD simulations using 152 water molecules, while the previously published AIMD simulation, which was reused here, was performed with 107 water molecules.[168]

For the ensemble generation, *NVT* simulations were performed. Here the constraint of the hydrogen bonds was removed and the timestep was reduced to 1 fs, while otherwise same settings were used. *NVT* simulations were performed for 40 ns. Snapshots for NMR calculations were taken every 0.1 ns resulting in 400 snapshots which were further used in NMR calculations.

For pressure-dependent FFMD simulations a larger solvent box was necessary to ensure the same cutoff without violating the minimum image convention even at high-pressure conditions. As no AIMD simulations were performed on the same systems technical limitations were not in place and the solutes were each placed in 278 water molecules. *NpT* simulations were performed at 1, 1000, 2000 and 3000 bar. Otherwise, the same settings as for the ambient pressure simulations were used. A small investigation of system size effect for DSS and NMA at ambient conditions showed no significant dependence of the resulting density on the system size justifying this approach and transferability from the smaller system simulations used for comparison to AIMD simulations.

To calculate the pressure-dependent magnetic susceptibility of water FFMD simulations were performed of pure water. Here 768 water molecules modelled by the TIP4P-2005[227] water model were simulated at 1 bar, 2 kbar and 10 kbar with a temperature of 298.15 K. One additional simulation at 1 bar and 293.15 K was performed. The cutoff was increased to 12 Å. Otherwise the same settings as before were used. The production run was performed for 60 ns using a time step of 1 fs. For the susceptibility calculation 400 snapshots were taken equidistantly from each simulation. Once again a benchmark against previous AIMD simulations at 1 bar and 10 kbar was performed.[235] A comparison of densities as one of the main underlying differences between FFMD and AIMD simulation will follow as part of the results in Table 8.

3.2.1 MLMD simulations using ANI-type Machine Learning Potentials

As a third simulation method MLMD simulations were performed using a variety of ANI-type MLPs described in Table 1.

Table 1: Models used in the MLMD simulations of NMA in *cis*- and *trans*-conformation.

Model	Description
ANI-2x	ANI-2x MLP from ref.[66]
ANI-dr	Modified ANI-2x MLP including dispersion[107] and repulsion[236] provided by I. Pickering.[237] The MLP was trained to the B97-3c level of theory.[238]
custom	Custom model provided by Y. Xue[239] based on the ANI-2x dataset with additional training data of alanine dipeptide-water clusters. The MLP was trained to the B97-3c level of theory.[238]

Simulations were performed in a development version of TorchANI-AMBER.[237] As the training data of these MLPs is limited to neutral molecules and the goal was to compare the method to AIMD and FFMD ensembles only NMA in *cis*- and *trans*-conformation was simulated using the MLPs. Here the Langevin thermostat[240] was used to keep the temperature at 298.15 K. Two systems with one solute and either 152 or 342 water molecules were simulated. A cutoff of 5.4 Å was used for AEV calculations. For production runs *NVT* simulations were performed for a simulation time of 1 ns from which 400 snapshots have been extracted for NMR calculations using either no explicit solvent or 4 Å of water molecules around all solute atoms together with TIP3P[241] point charges on all other solvent atoms.

3.3 Ensemble NMR calculations

All NMR calculations for the structural ensembles were performed at the OLYP/6-311+G(d,p)[110][111] level of theory with Gaussian16 rev.C01[242] using the Gauge-Independent Atomic Orbital (GIAO) method [219][220] to compute the nuclear magnetic shielding tensors. Selected calculations were performed on the MP2 level of theory with the same basis set which will be mentioned in the according chapters. For the calculation of chemical shifts, the difference between the isotropic shielding constants of DSS and the nucleus of interest was calculated, all averaged over magnetically equivalent nuclei. For calculations with the PCM the IEFPCM [117] was used with standard settings within Gaussian.[145]

All EC-RISM calculations were performed on a cubic grid with 120^3 points with 0.3 Å spacing in all spatial directions. Convergence criteria were set to 10^{-6} for the maximum direct correlation function deviation in 3D RISM calculations and a maximum free energy difference of 0.01 kcal mol^{-1} between two EC-RISM cycles. The 3D-RISM equations were solved using the MDIIS[243] method using four MDIIS vectors with a mixing factor of 0.4 with which previous iterations are included. Solute Lennard Jones parameters of TMAO for EC-RISM calculations were taken from TMAO-V3[244] while GAFF 1.4[121] parameters were used for all other solutes. For DSS silicon parameters from Makrodimitri *et al*.[245] were added following the previously described procedure for the DSS parametrization.[8] Explicit water molecules as well as EC-RISM water susceptibilities were modeled at 298.15 K using modified SPC/E charges and Lennard-Jones parameters in line with previous computations.[33] The exact parameters of this model are given in Table 2.

Table 2: Geometry and non-bonded parameters of the modified SPC/E water model in line with previous computations. [33] During hybrid calculations with explicit water molecules and EC-RISM these parameters were used for all explicit water molecules.

Atom	*x* / Å	*x* / Å	*x* / Å	*q* / e	*σ* / Å	*ε* / zJ
O	0.000	0.000	0.000	-0.8476	3.1660	1.0797
H	-0.815	0.000	0.579	0.4238	1.0000	0.3891
H	0.815	0.000	0.579	0.4238	1.0000	0.3891

Atom centered point charges were determined with the CHelpG scheme[246] employing default radii and used with the additional constraint to match the quantum chemically obtained dipole moment. Throughout all ensemble NMR calculations the PSE-2 closure was used.[170] A limited number (< 10%) of non-converged calculations depending on the solute and solvent system were disregarded.

Pressure-dependent NMR calculations including an EC-RISM solvation, either solely or in combination with explicit solvation, were performed using previously determined[8] solvent susceptibilities with

pressure-dependent densities, dielectric constants and compressibilities listed in Table 3. The solvent susceptibilities were calculated from 1D-DRISM using the HNC closure.

Table 3: Pressure-dependent relative dielectric permittivities ε, number densities ρ, and isothermal compressibilities κ of water used for the generation of solvent susceptibilities, calculated from the equation of state (EOS)[247] and, for κ, also from 1D-DRISM/HNC. Susceptibilities were taken from earlier work.[8]

p/bar	ε	ρ / Å^{-3}	κ_{EOS} / 10^9 Pa^{-1}	$\kappa_{DRISM/HNC}$ / 10^9 Pa^{-1}
1	78.4470	0.0333295	0.450183	0.714088
500	80.3705	0.0340416	0.399007	0.633388
1000	82.0662	0.0346904	0.357597	0.569452
2000	84.9467	0.0358344	0.295023	0.474971
3000	87.3377	0.0368204	0.250155	0.408612
10000	97.7471	0.0414170	0.116971	0.215218

To calculate the pressure-dependent magnetic susceptibility 50 snapshots from the AIMD simulation of water at 1 bar and 10 kbar were provided by D. Marx.[235] Together with the snapshots from FFMD simulations the calculation of the magnetic susceptibility was performed for a water sphere containing 67 water molecules. This number was chosen as close to the maximum size sphere inside the AIMD simulation box. For consistency this was repeated for the larger FFMD simulation cell. Magnetic susceptibility calculations were performed using additional EC-RISM solvation with the previously described settings. After the convergence of the EC-RISM cycle a single point NMR calculation was performed to calculate the magnetic susceptibility with Gaussian16 rev.C01[242] using the Gauge-Independent Atomic Orbital (GIAO) method.[219][220]

3.4 Pressure-dependent geometry optimizations using EC-RISM

For the molecules shown in Figure 6 pressure-dependent geometry optimizations were performed using EC-RISM. Here a development version of ORCA[199] provided by P. Kibies and M. A. Garcia-Rates with a direct EC-RISM implementation was used. An EC-RISM grid of 128^3 grid points with a 0.3 Å spacing was used for all geometry optimizations. Geometries were optimized using the B3LYP/6-311+G** [111][102] level of theory at pressures of 1, 500, 1000, 2000 and 3000 bar using very tight convergence criteria for the optimization and tight convergence criteria for the SCF cycle. The choice of very tight convergence criteria for the optimization was deemed to be necessary as the pressure-dependent change in bond lengths was found to be less than the default convergence criteria. Solvent susceptibilities were again taken from earlier work,[8] see Table 3.

3.5 pH dependency of the EPR spin probe HMI

To investigate the pH dependency of the EPR spin probe HMI previously established methodology used to calculate EPR parameters for the neutral from of HMI[9][248] was applied to the protonated form of HMI ($HMIH^+$). As two different diastereomers of $HMIH^+$ can potentially exist the Gibbs energy difference was calculated using an indirect approach. First, the solvation free energies were calculated with EC-RISM and the revPBE0-D3[112][113][107] functional with the def2-TZVPP[84] basis set with decontracted s-functions on CPCM-optimized structures. Second, after optimization in vacuum the differences of gas phase energies were determined at the same level of theory followed by single point DLPNO-CCSD(T)[249][250] calculations with the same basis set and added to the solvation free energies.

Based on AIMD simulations performed by B. Sharma a hierarchy of solvent models was used to disentangle structural from solvation effects. The models are summarized in Table 4.

Table 4: Solvent models used to calculate EPR parameters of HMI and HMIH+ with respective use of explicit solvation and additional background solvation.

Method	Explicit solvation	Background solvation
VD	-	-
SSS	5 Å around nitroxide-O	-
QM/MM	5 Å around nitroxide-O	TIP3P[241] point charges
EC-RISM	5 Å around nitroxide-O	EC-RISM

First a vertically desolvated ensemble (VD) was used where the solute was stripped of all solvent during the EPR calculation. Additionally an explicit second solvation shell of water molecules around the nitroxy oxygen (SSS) was added and two hybrid solvation models were used by addition of either EC-RISM or force field-based background charges (QM/MM) on top of SSS snapshots. Compared to the NMR calculation this follows the “local” explicit solvation approach with a radius of 5 Å around the nucleus of interest.

Subsequent EPR calculations were performed on the revPBE0-D3 [107][112][113] level of theory using ORCA version 5.0.3.[199] The def2-TZVPP basis set[84] with decontracted s-functions was used throughout all calculations. EPR calculations were also redone for HMI for which previously calculations were performed using the QM/MM and EC-RISM solvent model.[9] This was done to achieve consistency between the software used. Lennard Jones parameters of HMI and $HMIH^+$ were taken from the GAFF 1.4 force field.[121] EC-RISM calculations were done using the same settings as described in chapter 3.3 with two exceptions. The grid dimensions used were 120 grid points with a spacing of 0.5 Å between grid points. A finer (0.3 Å) grid spacing was briefly investigated but showed no significant difference. A small subset of calculations was also performed using 46 grid points with a spacing of 0.5 Å between grid points. In addition the PSE-3 closure was used in line with earlier work.[9]

The GIAO[219][220] method was used to determine all EPR parameters, namely the *A*- and *g*-tensor from which isotropic quantities (g_{iso}, A_{iso}) were taken as one third of the respective trace in addition to individual tensor elements which were also analyzed. All EPR calculations were performed for the same set of 1000 snapshots for HMI and $HMIH^+$ respectively, taken from AIMD simulations.

3.6 Training of a solvated ANI-type Machine Learning Potential

From the original ANI-2x dataset[66] structures were taken equidistantly to maximize the diversity of the training dataset. Due to the high computational demand of the EC-RISM calculation only every 20th structure of the original dataset was considered. The exact number of structures for each number of atoms is shown later together with resulting observables in Figure 33.

For all structures EC-RISM calculations were performed using the r^{2}SCAN-3c[251] level of theory. Lennard-Jones parameters were taken from the GAFF 1.4 force field.[121] Structures containing bonds outside of the default bond lengths of the antechamber atom typer[252] were disregarded. EC-RISM calculations were performed using the ORCA quantum chemistry program version 5.0.3.[199] EC-RISM calculations used the settings from chapter 3.3 with the exception that ten MDIIS vectors were used. The solvent susceptibility for 1 bar and 298.15 K was taken from Table 3. The partial molar volume was taken from the direct correlation function[253] and calculated using the 1D-RISM estimate of the isothermal compressibility of $0.714088 \cdot 10^9$ Pa^{-1}. Non-converging calculations were disregarded afterwards. In total 380575 datapoints were generated.

To normalize observables such as G^{sol} ground state atomic free energies in solution, still referred to as GSAE consistent with other ANI-type MLPs, were calculated for each element contained in the ANI-2x dataset. The results are shown in Table 5.

Table 5: Ground state atomic energies used for training of the ANI-EC-RISM dataset. The values were calculated at the r^{2}SCAN-3c level of theory.

Atom	GSAE / E_h
H	-0.490920596
C	-37.82318106
N	-54.57152480
O	-75.21043138
F	-99.71773388
S	-398.0725309
Cl	-460.1051849

For training of the MLP the workflow for the training of earlier ANI-type models was followed.[10][66] This includes the architecture of the MLP. As for the ANI-2x model a radial cutoff of 5.1 Å was used,

while the angular cutoff was set to 3.5 Å. The radial AEVs were constructed from 16 differently shifted symmetry functions while angular AEVs used eight. This results in a total of length of 1008 for the AEVs. The architecture of the neural networks was taken from the ANI-2x model and is illustrated in Table 6.

Table 6: Dimension of the three hidden layers for each species of the ANI-EC-RISM MLP. The dimensions were taken from the ANI-2x model.[66]

Atom	Dimension of hidden layers
H	256, 192, 160
C	224, 192, 160
N	192, 160, 128
O	192, 160, 128
F	160, 128, 96
S	160, 128, 96
Cl	160, 128, 96

All hidden layer nodes used a Gaussian activation function[254] while the output node had a linear activation function. This allowed for the automatic differentiation of the energetic output of the MLP with respect to the input layer, which is the AEV. From this differentiation forces can be calculate and used for geometry optimizations, which were done using the auto3d program.[77]

All data was split with a ratio of 80/20 into training and validation data. The training was performed within a custom training environment provided by K.K. Huddleston[255] using the Torchani[256] program. The models were trained to individual energetic outputs of an EC-RISM calculation which will be shown later in Table 35 together with validation RMSE values in Table 36. Datapoints were batched with a size of 500. The network weights were optimized using the adamW update method.[257][258] For the adamW algorithm an improvement in RMSE of 0.0006 kcal/mol was deemed an improvement which should be reached after a maximum of 14 epochs. If not, the learning rate was multiplied by a factor of 0.7. If a learning rate of 10^{-7} was reached using these hyperparameters the training was stopped. Using this setup ca. 700 epochs of training were performed until early termination while ca. the last 200 epochs were performed without an improvement.

4. Results

4.1 Model calculations to determine structural sensitivity of NMR parameters

The goal of this chapter is to demonstrate the structural sensitivity of NMR parameters and identify key structural parameters which influence calculated NMR parameters. This is done to inform other investigations shown in this thesis. First, molecular parameters such as bond lengths, both intra- and intermolecular, and dihedral angles are investigated for model systems. In addition, the influence of high hydrostatic pressure is investigated, disentangling both the influence of structural and solvation changes on the chemical shift changes under high pressure conditions. This adds to previous calculations which only included the change in solvation.[7]

Lastly the pressure-dependent magnetic susceptibility is investigated validating an experimentally found increase of the mass susceptibility of water. Here ensemble-based calculations from AIMD and FFMD simulations are used in line with previous studies of the magnetic susceptibility at ambient conditions which relied on AIMD simulations.[259]

4.1.1 Model calculations of chemical shifts with respect to structural parameters

NMR parameters are known to be structurally sensitive.[260] Any chemical shift prediction, as will be performed in this thesis, requires an accurate representation of the underlying structure, regardless of the system.

To outline this sensitivity for chemical shifts a scan of NMR parameters was done based on a variety of structural parameters. These include the N-H bond length of *trans*-NMA (Figure 7, **A**) and an explicit hydrogen bond length of NMA with water (Figure 7, **B**). This will set the stage and be used as benchmarks for further NMR calculations where the used methods will result in different intramolecular parameters as well as intermolecular distances of hydrogen bonds. The model system of NMA, resembling the protein backbone, will be used for all further studies as well.

In addition the dihedral angle of propionic acid (Figure 8) will be analyzed. Here propionic acid on the one hand is a model for the sidechain glutamate of a peptide, which was studied both experimentally[261] and theoretically[262] in terms of NMR parameters at different pH values, on the other hand the analysis of a dihedral angle and the dependence on the chemical shift will also be performed for other systems.

These model calculations are done to provide a benchmark for further calculations to disentangle possible error sources by how much they vary in structural parameters and how this lines up with errors in NMR parameters.

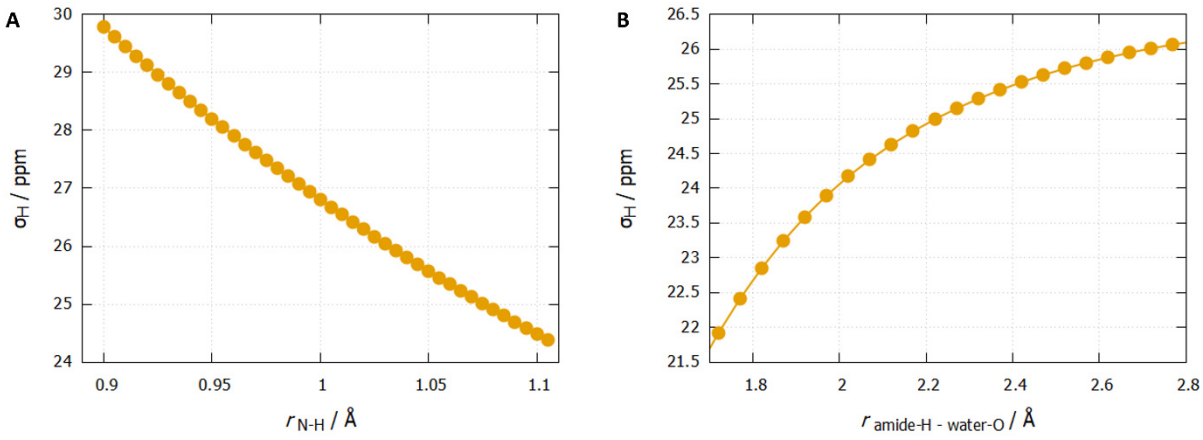

Figure 7: Nuclear shielding constants of the amide proton of trans-NMA in dependence of the N-H bond length (**A**) and the hydrogen bond length to an explicit water (**B**) for both calculations the OLYP/6-311+G**/EC-RISM/PSE-2 level of theory was used. For the explicit water calculation a configuration from FFMD-simulation was taken with a distance of 2.17 Å between the amide proton and the water oxygen.

For NMA a change of -0.3 ppm/pm is found for the N-H bond length. This is consistent for the whole range from 0.9 to 1.1 Å.. For the distance to an explicit water a nonlinear effect is found. For closer distances of up to 2 Å the shielding constant increases by ca. 0.1 ppm/pm while for distances of more than 2.2 Å this rate decreases to ca. 0.03 ppm/pm.

Figure 8 shows a clear difference between protonated and deprotonated propionic acid. In terms of the energetic profile the protonated propionic acid shows one significant minimum favored by more than 1 kcal/mol. In terms of shielding constants, the energetic minimum corresponds to the highest shielding constant found for the protonated propionic acid. This is also true for the deprotonated species, however here two distinct minima are found due to the symmetric carboxylate. These then share the same highest shielding values which are more than 3 ppm lower than the maximum found for the protonated form.

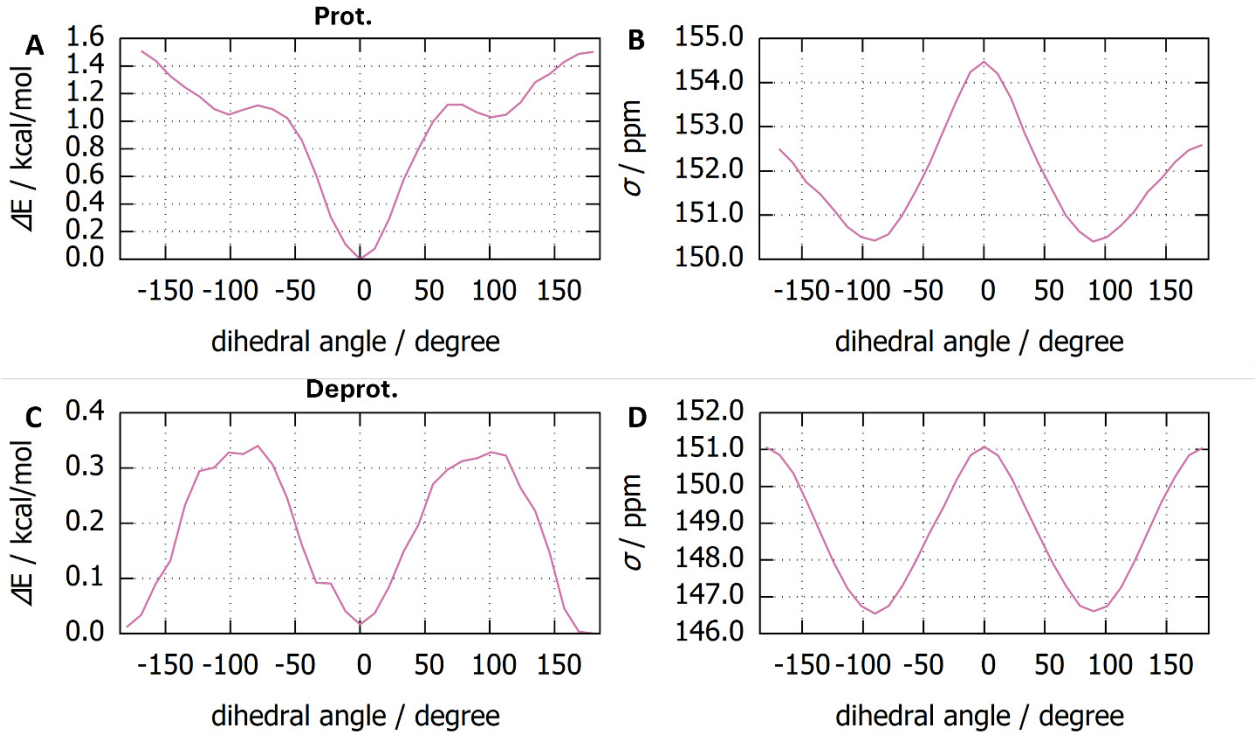

Figure 8: Dihedral scan of propionic acid with a protonated (**A, B**) and deprotonated (**C**,**D**) carbonyl group. It shows the potential variation of the C_{β}- C_{γ}- $C_{carbonyl}$-O dihedral angle. Panels **B** and **D** show the shielding constants of the C_{γ} for this aspartic acid. Calculations were done on the OYLP/6-311+G**/EC-RISM(water)/PSE-2//B3LYP/6-311+G**/CPCM(water) level of theory.

Looking at the overall NMR profile shows that already for a simple small molecule significant chemical shift changes can be found in a relatively tiny energetic range around the local minimum. It is expected that purely the thermal energy and random molecular motion at ambient conditions is enough to vary the dihedral angle. This would then result in a significant change in the chemical shift. For example a dihedral angle of 30° would increase the chemical shift by ca. 1.5 ppm. A method such as MD simulations would be necessary to sample the local minima of these compounds accordingly to get accurate absolute chemical shifts. However, if only trends with respect to the protonation pattern are of interest, single point calculations might also be sufficient as some error cancellation can be expected, as both the protonated and deprotonated forms show similar behavior close to the local minima.

4.1.2 Disentangling of structural and solvation response to high hydrostatic pressure

In general, pressure can affect two different aspects of an NMR prediction. First, the structure can change due to the influence of high pressure. Second, during the NMR calculation high pressure can be applied by changing the solvent susceptibility within EC-RISM, resembling high pressure conditions. Previously using EC-RISM, it was only possible to account for the second part of change in the electronic structure under high pressure but using a fixed geometry optimized at ambient conditons.

One main tool to do this is with EC-RISM geometry optimization which were newly implemented within the ORCA quantum chemistry software. Based on EC-RISM disentangling structural and solvation effects is possible. To disentangle the effects, it was systematically varied when pressure was applied for the example of the amide proton of *cis*-NMA. A schematic representation of this is given in Figure 9.

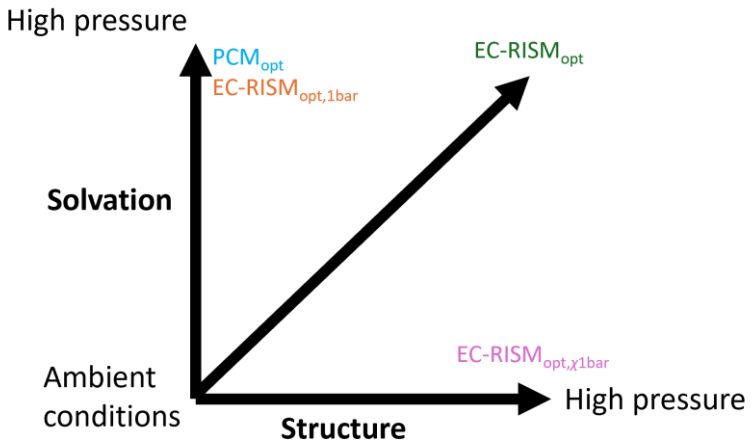

Figure 9: Schematic representation of the different methods in how they apply ambient conditions or high pressure in terms of the structural and solvation effect. The color scheme is the same as in Figure 10.

The results are given in Figure 10. The previously used method, called "PCM$_{opt}$", where pressure is applied only in the NMR calculations is similar to "EC-RISM$_{opt,1bar}$", but only uses a different method for the ambient-condition optimization, namely PCM instead of EC-RISM. These methods show the solvation part of the pressure dependence. In contrast "EC-RISM$_{opt,\chi 1bar}$" takes pressure-dependent geometries and uses ambient pressure for the NMR calculations. This method shows the structural part of the pressure dependence. "EC-RISM$_{opt}$" denotes the full pressure-dependent approach where both parts are accounted for. This can also be reconstructed from the individual calculations as "opt,1bar+opt, χ1bar".

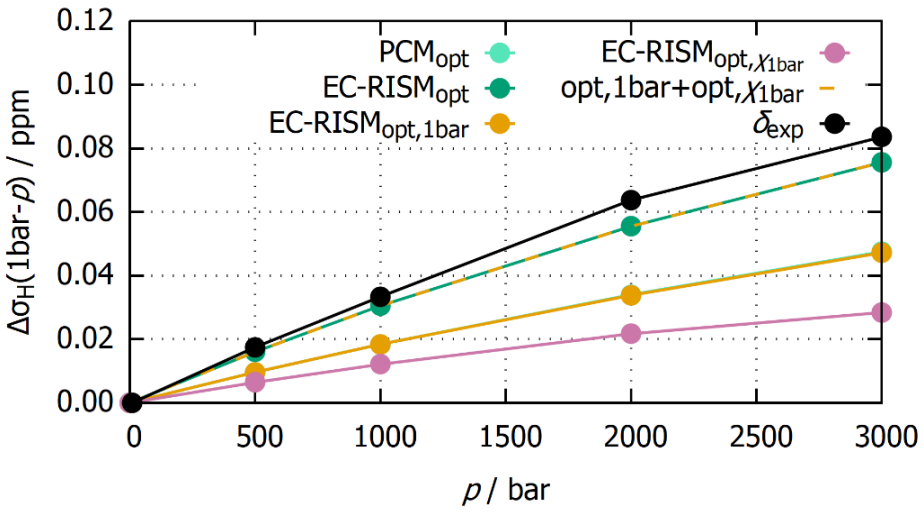

Figure 10: Changes in the nuclear magnetic shielding constant of the amide proton of NMA in the *cis*-conformation. The results were calculated using the OLYP/6-311+G** level of theory. EC-RISM$_{opt}$ denotes that EC-RISM was used for both the pressure-dependent geometry optimization and NMR calculation. For EC-RISM$_{opt,1bar}$ the geometry was optimized at 1bar and used for pressure-dependent NMR calculations. This approach was also followed for PCM$_{opt}$ where PCM was used for the geometry optimization, this is the previously used method.[7][8] For EC- RISM$_{opt,\chi 1bar}$ the inverse approach was used with pressure-dependent geometries using EC-RISM and NMR calculations at 1 bar. For opt,1bar+opt, χ1bar both individual approaches were added up. Every method is compared to experimental chemical shift differences. The difference in shielding constants was calculated in a way that is directly comparable with chemical shift differences if a pressure independent reference is assumed.

As a result for the example of *cis*-NMA the solvent change accounts for 2/3 of the overall pressure dependence (”EC-RISM$_{opt,1bar}$”) and the structural change accounts for the remaining 1/3(“EC-RISM$_{opt,\chi 1bar}$”). Only by including both effects the experimental trend can be matched quantitively. Here however it does not matter how the ambient condition structure is generated as “PCM$_{opt}$” and ”EC-RISM$_{opt,1bar}$” overlap. The disentangling in these two parts work quantitatively as the addition of individual parts (“opt,1bar+opt, χ1bar”) perfectly matches the direct calculation which includes both factors “EC-RISM$_{opt}$”.

Furthermore, upon investigating the structural changes, an elongation of the N-H bond with increasing pressure can be found. This corresponds to the increase in chemical shift following the basic analysis shown in Figure 7. The pressure-dependent bond length is shown in Table 7.

Table 7: Pressure-dependent N-H bond length of *cis*-NMA from geometry optimizations using EC-RISM at the B3LYP/6-311+G** level of theory. The ambient condition value using PCM-optimized geometries at the same level of theory is also shown as a reference.

p / bar	$r_{N\text{-}h}$ / Å
1	1.0151
500	1.0152
1000	1.0153
2000	1.0154
3000	1.0155
1 (PCM)	1.0105

Even if the change in bond length is small the pressure induced response in NMR parameters is so large that even these minor changes in the geometry can account for a significant portion of the NMR response and the inclusion in the calculations is necessary for quantitative agreement as shown in Figure 10. Taking the model calculation of Figure 7, as a reference an increase of 0.04 pm which was found at 3000 bar contributes ca 0.012 ppm to the amide chemical shift, roughly half of the structural part of the chemical shift increase shown in Figure 10 ("EC- RISM$_{opt,\chi 1bar}$"). The small increase in bond length is atypical as, especially in gas phase, high pressure leads usually to a contraction in bond lengths.[5]

It is noteworthy that the small changes in equilibrium bond length are much smaller than the vibrational changes at room temperature. Ensemble based methods like MD simulations will be investigated further in the next chapters where the NMR response to the change in average bond length will also be analyzed. This will further show if even small changes in the equilibrium geometry can be analyzed reliably or if they are overcompensated by the molecular vibrations.

To further investigate the change in structural parameters the change in bond length for the high-pressure dataset shown in Figure 6 is shown in Figure 11.

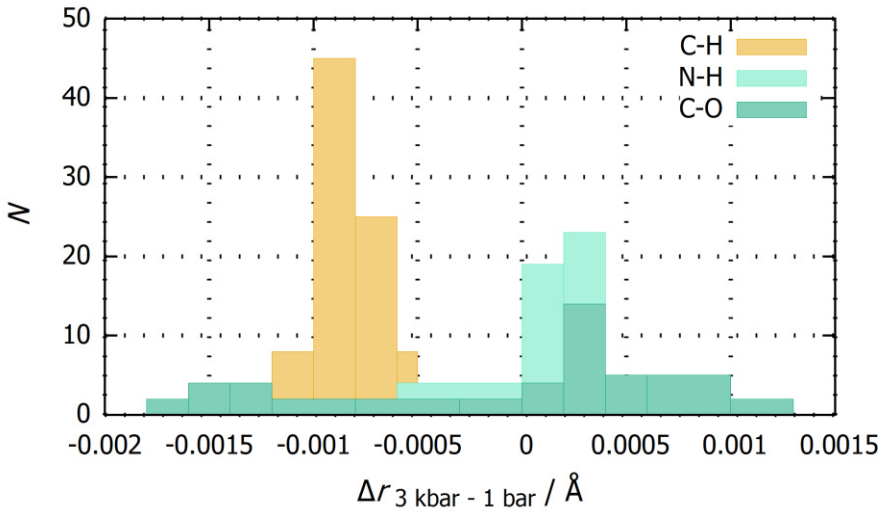

Figure 11: Histogram of changes in bond length for molecules of the high-pressure dataset (Figure 6) from 1 bar to 3 kbar calculated using the B3LYP/6-311+G**/EC-RISM level of theory. Bonds are divided into C-H (orange) N-H (blue) and C-O (green). Raw data is published in the electronic supplementary material, see chapter 6.1.

Figure 11 shows that the overall trend is a mixed trend of the bond-length change under pressure. For the N-H bonds a small increase in bond length can be seen consistent with the previous example of *cis*-NMA. This also shows that for hydrophobic parts like the C-H the gas phase trend of a decrease in bond length can be translated to solution with a consistent decrease of ca. 0.1 pm at 3 kbar. For the C-O bond no clear trend can be seen and both an increase and a decrease can be found. For the polar N-O bond of TMAO AIMD simulations showed an increase in bond length at 10 kbar[168] in line with the findings for the N-H and parts of the C-O bonds shown here. An increase in bond length was also found for some transition-state bonds using the XP-PCM[154] showing that this mixed response as seen here is plausible.

Overall however, it is noteworthy that including even these small structural changes is important to quantitatively match chemical shift changes under pressure. A deeper analysis of this will follow in chapter 4.4.

4.1.3 Pressure-dependent calculations of the magnetic susceptibility

As another NMR observable the magnetic mass susceptibility of water was calculated for high-pressure conditions. The goal was to validate an experimentally found trend of an increasing mass susceptibility. Furthermore theoretical methods are compared in terms of their accuracy and reasons for deviations are investigated.

For a direct comparison the same observable needs to be compared which in this case is the mass susceptibility χ_M. Experimentally however, the volume susceptibility χ is measured but it can be converted to the mass susceptibility by

$$\chi_M = \frac{\chi}{\rho}. \tag{99}$$

The output of quantum chemistry calculation is the susceptibility defined by equation (92). The output from the quantum chemistry program Gaussian[242] χ_{calc} in cgs-ppm can be converted to the mass susceptibility by

$$\chi_M = \frac{\chi_{calc}}{4\pi M_{tot}} 10^{-6} \tag{100}$$

using the total molar mass M_{tot} of the system which here was calculated with M_H = 1.00783 g/mol and M_O = 15.99491 g/mol.

Within this study AIMD and FFMD simulations were used as a basis to calculate the magnetic susceptibility for a sphere of 67 water molecules. Explicit water molecules were embedded within an EC-RISM background solvation. Snapshots of the AIMD simulations were provided by D. Marx.[235] Pressure-dependent experimental data of the magnetic susceptibility was provided by H.R. Kalbitzer.[263] As the AIMD was performed at 300 K compared to experimental data at 293.15 K and FFMD at 298.15 K another FFMD was performed at 293.15 K to estimate the error introduced by the change in temperature. The data is summarized in Figure 12. As one of the main sources of error and also to further investigate the underlying reasons for the pressure dependency of the mass susceptibility the influence of the density on the mass susceptibility was investigated and is shown in Figure 12, **B**. The pressure-dependent densities used are shown in Table 8.

Table 8: Pressure-dependent densities from the FFMD and AIMD. It is noteworthy that the AIMD was performed at 300 K while the FFMD was performed at 298.15 K.

p / bar	ρ FFMD / kg/m^3	ρ AIMD/ kg/m^3
1	1007.35	996.32
2000	1084.22	-
10000	1254.72	1237.10

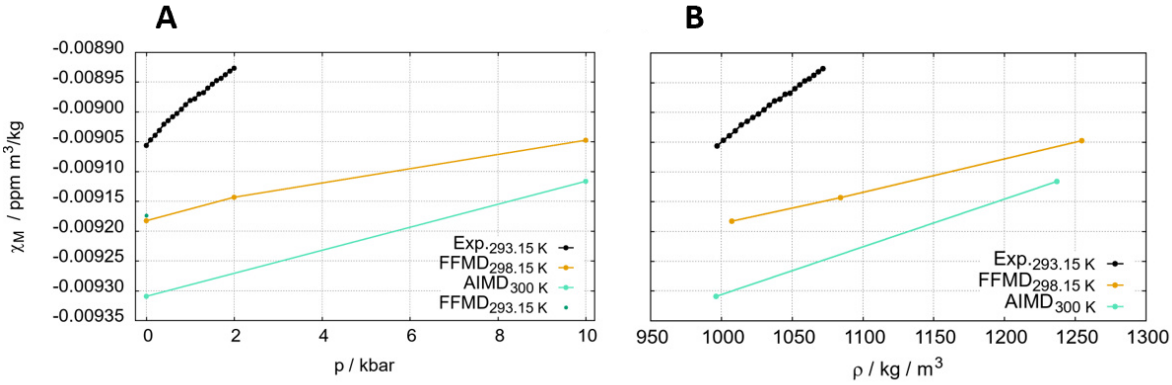

Figure 12: Pressure-dependent mass susceptibility (using the SI unit definition) of water with respect to the pressure (**A**) and density (**B**). Data is from experiments (black), based on AIMD snapshots at 300 K (blue) and FFMD snapshots (orange) at 293.15 K (triangles) and 298.15 K (points). Snapshots of the AIMD simulations were provided by D. Marx.[235] Experimental data was provided by H.R. Kalbitzer.[263] Calculations were performed on the OLYP/6-311+G**/EC-RISM level of theory.

Figure 12, **A** shows that the calculations match well with the experimental data. In terms of absolute values errors at ambient conditions are less than 3% in line with previous theoretical investigations[259] and the trend of an increase in mass susceptibility is matched for both the AIMD and the FFMD. Here the AIMD, even though the absolute values are further away from experiments, better matches the pressure trend that is further underestimated by the FFMD.

Figure 12, **B** shows a linear dependency of the mass susceptibility with respect to density which was first identified by S. Kast. This is true for experiments and calculations. Thus, the nonlinear part of the pressure dependent mass susceptibility as shown in Figure 12, **A** can be attributed to the change in density under pressure. This holds true regardless of method. The calculations however show a smaller of the mass susceptibility with respect to the density. The slopes of $5 \cdot 10^{-7}$ for FFMD and $8 \cdot 10^{-7}$ for AIMD compared to $19 \cdot 10^{-7}$ from experiments are hinting at the fact that further factors, like the molecular structure might play a role, these are expected to be more accurate from the AIMD in line with the higher slope found here.

In conclusion the theoretical workflow was suitable to reproduce the experimental pressure-dependent mass susceptibility. This is true for both the AIMD and FFMD snapshots. Overall, an increase in mass susceptibility was found with an increase in pressure. The magnitude of this trend gets consistently underestimated with the AIMD more closely matching it compared to the FFMD. An absolute error of 2-3 % remains even at ambient conditions, while the FFMD more closely matches the absolute values. This error can most likely be attributed to the chosen level of theory. As the reason for the nonlinear pressure-dependent change in the mass susceptibility the change in density was determined for which a linear correlation was found. However, this investigation also revealed an additional dependence beyond the change in density, which is found both experimentally and theoretically and might be linked to the molecular structure of the water.

4.1.4 Summary and outlook

In summary these exemplary calculations highlighted the structural sensitivity of NMR parameters. Small errors in the solute geometry or the local solvation determined by theory would show up in significant errors of the chemical shift. This is regardless of the level of theory used in NMR calculations which was not varied for these exemplary calculations.

Investigation of a dihedral scan of propionic acid highlighted the fact that also local conformational minima need to be sampled accurately to avoid significant errors. Here methods such as MD simulations are preferred and will be extensively analyzed in the next chapters of this thesis. These exemplary calculations will act as reference data to understand correlations between structural changes and resulting changes in NMR parameters. Structural parameters that will be analyzed include intramolecular bond lengths, hydrogen bond lengths based on solvent distributions and dihedral angles. All have been shown to drastically change predicted chemical shifts.

An investigation of pressure-dependent chemical shift changes of *cis*-NMA revealed that 2/3 of the chemical shift change can be attributed to change in solvation while 1/3 can be attributed to the structural change. Even though the bond length of the N-H bond changes only ca. 0.04 pm at 3000 bar, which was confirmed for more molecules, these small changes need to be included to quantitatively match the small experimental chemical shift changes.

Lastly, the pressure-dependent magnetic susceptibility of water was investigated at high-pressure conditions using both FFMD and AIMD ensembles. At ambient conditions the errors were around 3 % with respect to experimental values in line with other theoretical investigations[259] which makes both methods a reasonable choice for these calculations. The experimentally found increase of the mass susceptibility was matched by both methods, however the magnitude of this trend was underestimated both times, with the AIMD still showing the higher increase compared to the FFMD. As one of the main reasons for the increase, a linear correlation with the density was found which increases under pressure. Another linear increase in addition to the increasing density was found which can be due to further intramolecular changes under high pressure, however this phenomenon could not be fully explained and requires further investigation.

4.2 The accuracy limit of chemical shift predictions for species in aqueous solution

The goal of this chapter is to analyze a variety of methods for the ensemble generation for NMR calculations. The general goal of this study is twofold. The first goal is to reach the highest level of accuracy in the calculation of a chemical shift with the main focus on an accurate description of the solvent, which is done by using explicit solvent, sampled from state-of-the-art AIMD simulations, together with additional background solvation for which the PCM, EC-RISM and a QM/MM-type model are used. Here explicit solvation is expected to yield the highest accuracy, which is why convergence studies with respect to the explicit solvent size will be performed. If the NMR parameter converges with respect to the explicit solvent size, it is considered accurate. The second goal is to identify sweet spots, which are computationally less demanding but yield only small errors with respect to the fully converged description of the solvent. Here an additional scheme for explicit solvation is introduced which only locally uses explicit solvation reducing the number of explicit solvent molecules and therefore computational cost.

The central quantity of this study is the chemical shift, which results from calculated nuclear magnetic shielding constants with respect to a reference. This is why, as a reference, DSS is analyzed. As a target molecule NMA, both in *cis*- and *trans*-conformation, is used again. For these molecules new AIMD simulations were performed. In addition, earlier AIMD simulations for the target molecule TMAO[168] are reused for a more general analysis. With this approach both the shielding constants as the primary information from calculations but also the chemical shifts, as the main experimental NMR parameter are analyzed. The choice of both *cis*- and *trans*-NMA allows to investigate methods in terms of how accurately they can distinguish between conformers, a possible application of developed methods in the context of structure elucidation. Both *cis*- and *trans*-NMA can be distinguished for all nuclei during experiments.

Overall, the chosen model compounds represent both polar and apolar solute-water interactions making them ideal benchmark candidates.

It is noteworthy that the level of theory for the NMR calculation is only varied at one point. For most calculations a relatively cheap level of theory, OLYP/6-311+G(d,p), is chosen to allow for this computationally demanding systematic analysis. Remaining errors in comparison with respect to experimental data are expected and can, most likely, be attributed to the level of theory.

In an attempt to further reduce the computational cost FFMD and MLMD simulations are used and compared to AIMD simulations, which remained the computational bottleneck for this study. With the analysis of structural parameters done before in chapter 4.1, the individual structural ensembles are analyzed with respect to the same structural parameters. This allows not only the identification but also the quantification of potential error sources.

This project extends on a previous master's thesis[264] where only FFMD simulations were used to calculate nuclear magnetic shielding constants of both DSS and TMAO with an increasing number of water molecules together with PCM and EC-RISM background solvation. During this master's thesis only a local explicit solvation approach was used. The comparison of the local solvation approach to a full solvation approach explicitly solvating the whole molecule is newly done as part of this thesis. Newly performed AIMD simulations of DSS now also allow for the calculation of chemical shifts from AIMD ensembles which was previously only available for TMAO. To overcome size differences between molecules, especially relevant for the newly treated NMA, a distance criterium is used within this thesis to ensure a similar solvation surrounding different nuclei. This is in contrast to using the same number of explicit solvent molecules to calculate chemical shifts as done in the master's thesis. Based on the distance criterion and to ensure comparability between different molecules FFMD simulations of DSS and TMAO and all results shown here were newly calculated as part of this thesis.

To analyze the effect of running calculations for an ensemble of structures from AIMD simulations first the size of the ensemble examined must be determined. To reduce the computational demand the goal is to use as little snapshots as possible while still capturing the conformational flexibility of the solute and solvent and therefore nonlinear effects on the NMR parameters, as seen before in chapter 4.1. For this investigation, an increasing number of snapshots was taken from AIMD simulations of DSS, NMA and TMAO (performed by B. Sharma).[265] The cumulative average of the isotropic shielding constants of selected nuclei of DSS, NMA and TMAO was calculated and the results are shown in Figure 13.

Figure 13 clearly shows convergence for the running averages. This shows that even a simulation time of 200 ps is sufficient to sample the depicted conformers. Different conformers that are proposed for DSS[8] can be simulated on the same short time scale by starting in a different initial conformer. Within the 200 ps of simulation time 400 snapshots resulting in a rate of 0.5 snapshots/ps are a sufficient sampling rate to reduce statistical errors to less than 0.1 ppm for ^{1}H and 0.2 ppm for ^{13}C. A lower sampling rate can be used if multiple equivalent nuclei are present as is the case for DSS or TMAO. For most of the simulations 200 snapshots might also be sufficient as no clear trend is visible afterwards except for the ^{13}C of TMAO where a small increase is even visible from 350 to 400 snapshots.

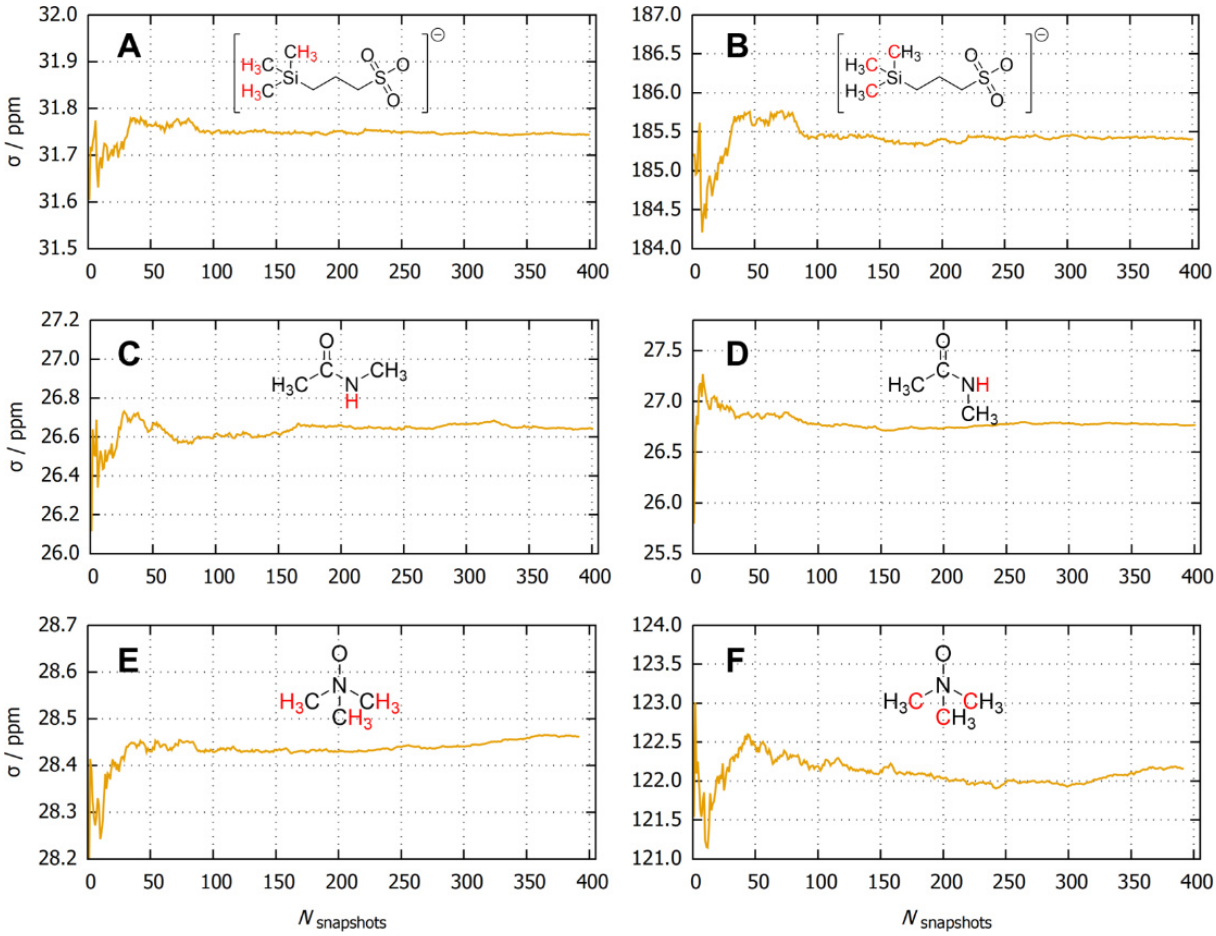

Figure 13: Cumulative average of the Isotropic shielding constants of the highlighted nuclei with increasing number of snapshots considered. The shielding constants were calculated using the OLYP/6-311+G(d,p) level of theory as an average over all magnetically equivalent nuclei per snapshot and afterwards averaged up to $N_{snapshots}$. Snapshots were taken as a time series from AIMD simulations. No solvation was considered for these calculations.

4.2.1 The solvent effect on ensemble calculations of NMR parameters

To investigate the solvent effect on NMR parameters based on AIMD simulations a multitude of methods to represent the solvent have been used. The goal of this investigation was to identify sweet spots in terms of accuracy and computational demand. The premise of this study was that explicit solvation should yield the highest accuracy but is also the most time consuming. Identifying sweet spots than relies on minimizing explicit solvation with potentially adding background solvation using an EC-RISM, PCM or QM/MM background solvation.

For the variation of explicit solvent two different selection criteria were used. A “local” approach where the solvent molecules were only taken around the nuclei of interest. This is especially interesting for NMR or EPR spectroscopy as the local environment is investigated and only a limited

number of nuclei might be of interest and show a significant signal in experiments for the underlying questions that are investigated.

In contrast to the “local” approach a “full” approach was used. Here solvent molecules are included all around the solute molecule. This should include solvent effects with the highest sophistication possible as a maximum number of molecules is included in QM calculations. The size of the QM region is then systematically varied based on a distance criterium up to 4 Å around the molecule. For DSS, as no convergence was initially visible the “local” solvation approach was taken even further up to 5.5 Å. An overview of the average number of water molecules for each solvent method is provided in Table 9.

To decrease the computational demand hybrid solvation systems have been set up using either EC-RISM or PCM as a background solvation together with explicit solvation. The goal is to identify a sweet spot where all solvent effects are included in the QM calculation while the computational demand is reduced as much as possible. The resulting shielding constants with an increasing number of explicit solvent molecules is shown in Figure 14. The points correspond to 0 - 5.5 Å of the explicit solvent region for all methods shown. However, more than 4 Å were only used for the “local” solvation of DSS.

Table 9: Average number of water molecules included in the NMR calculations of the molecules of interest for a given radius using the “local” and “full” solvation method. Ensembles were generated from AIMD simulations performed by B. Sharma.

r_{water}	DSS_{local}	DSS_{full}	*trans*-NMA_{local}	*trans*-NMA_{full}	*cis*-NMA_{local}	*cis*-NMA_{full}	$TMAO_{local}$	$TMAO_{full}$
-	0	0	0	0	0	0	0	-
2.5 Å	5.4	13.3	1.1	6.7	1.3	6.5	5.7	7.1
3.0 Å	14.5	25.4	2.1	14.8	2.9	14.7	14.9	14.9
4.0 Å	27.9	43.3	7	28.3	8.2	27.9	25	28.3
4.5 Å	35.2	-	-	-	-	-	-	-
5.0 Å	44.9	-	-	-	-	-	-	-
5.5 Å	56.9	-	-	-	-	-	-	-

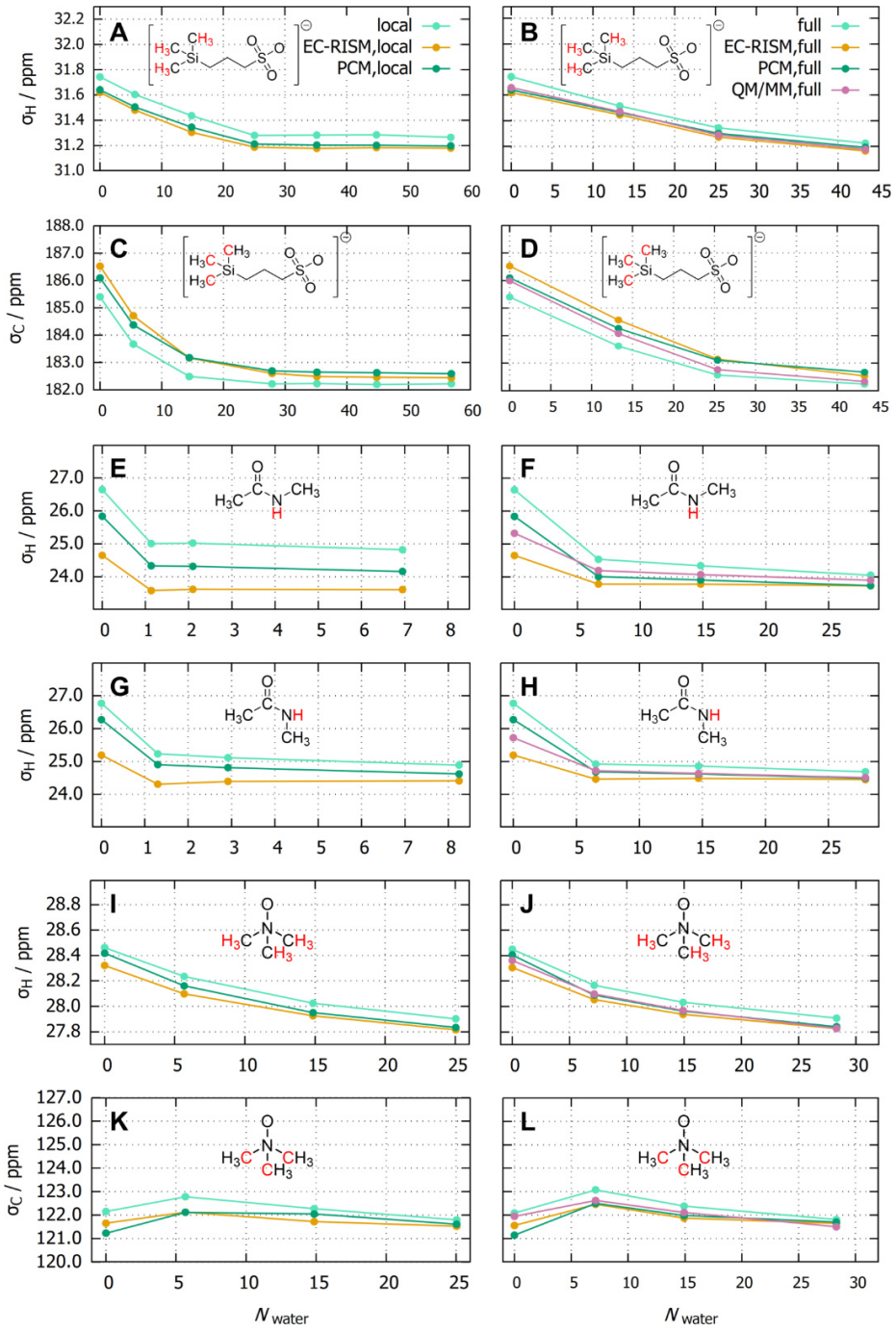

Figure 14: Isotropic shielding constants of the highlighted nuclei of the highlighted nuclei with increasing number of explicit water molecules included in the NMR calculations. The OLYP/6-311+G(d,p) level of theory was used for all calculations. The "local" solvation (left column) denotes that the water molecules were selected only around the highlighted nuclei while "full" (right column) denotes that the water molecules were included around the whole molecule. Shielding constants were averaged over snapshots and all magnetically equivalent nuclei. The EC-RISM and PCM subscripts indicate that an additional solvent background modelled with EC-RISM and PCM was used while "QM/MM" denotes that additional TIP3P-point charges[241] were placed on the water molecules outside of the explicit QM zone. The last point of the "QM/MM,full" curve is proposed as the computational reference. It considers the whole simulation box either explicitly up to 4 Å around the whole molecule or as point charges. Raw data is published in the electronic supplementary material, see chapter 6.1.

Depending on the specific nucleus convergence is reached at different explicit solvent shell sizes. For the amide protons of NMA even 2.5 Å of "local" solvation (ca. 1 explicit solvent molecule) is enough to reach convergence while for DSS 4 Å of explicit solvation are necessary to capture all solvent effects. This is referenced against the last point of the "QM/MM,full" curve which includes the whole simulation point at least as point charges with the largest feasible explicit solvation shell and is therefore further referenced as "$ref_{QM\,/\,MM}$".

In the comparison of the different background solvation models the "local" solvation is especially interesting as this approach already reduces the number of solvent molecules and gets impacted most by the background solvation. This is especially visible for NMA (Figure 14: **E**, **G**) where a difference of more than 1 ppm remains even with 4 Å of explicit solvation between the use of EC-RISM and no background solvation. This shows that accurate solvation of the carbonyl group is important for the amide chemical shift (Figure 14: **E**, **G**) and EC-RISM yields an accurate background solvation in comparison to PCM. For PCM almost 0.5 ppm difference remain compared to the "$ref_{QM\,/\,MM}$" value (Figure 14: **F, H**) where the carbonyl group is surrounded by explicit water molecules.

To further summarize the findings, an overview of nuclear shielding constants for selected points together with the single structure approach denoted as "opt" is given in Table 10. The "local" solvation approach here corresponds to 2.5 Å of "local" solvation while "full" corresponds to 4 Å of "full" solvation. These points are significant as they include the minimum amount of explicit solvation as well as the converged explicit solvation. A "0" solvation approach is also given where no explicit solvation is used.

Table 10: Isotropic shielding constants / ppm for all highlighted nuclei, averaged over magnetically equivalent atoms, for a 4 Å "full" (AIMD ensemble) explicit solvation shell augmented by an EC-RISM ("ECR"), PCM, and QM/MM (corresponding to "$ref_{QM/MM}$" in the figures, relative to which all other numbers are reported) background and corresponding 2.5 Å "local" data; EC-RISM and PCM results for AIMD ensembles with 0 explicit water molecules (subscript "0") and for optimized structures from OLYP/6-311+G(d,p)/PCM (superscript "opt", subscript "0"). Raw data is published in the electronic supplementary material, see chapter 6.1.

Nucleus	$\sigma^{AIMD}_{QM/MM,full}$	$\Delta\sigma^{AIMD}_{QM/MM,0}$	$\Delta\sigma^{AIMD}_{ECR,full}$	$\Delta\sigma^{AIMD}_{ECR,local}$	$\Delta\sigma^{AIMD}_{ECR,0}$	$\Delta\sigma^{opt}_{ECR,0}$	$\Delta\sigma^{AIMD}_{PCM,full}$	$\Delta\sigma^{AIMD}_{PCM,local}$	$\Delta\sigma^{AIMD}_{PCM,0}$	$\Delta\sigma^{opt}_{PCM,0}$
DSS_H	31.18	0.48	-0.02	0.30	0.44	0.56	0.01	0.32	0.46	0.59
DSS_C	182.33	3.66	0.21	2.38	4.20	5.09	0.34	2.04	3.77	4.80
cis-NMA_{NH}	24.51	1.21	-0.06	-0.20	0.68	1.16	-0.02	0.39	1.76	2.22
trans-NMA_{NH}	23.90	1.42	-0.17	-0.32	0.75	1.14	-0.17	0.43	1.94	2.25
$TMAO_H$	27.83	0.53	0.00	0.27	0.49	0.72	0.01	0.32	0.59	0.82
$TMAO_C$	121.50	0.44	0.14	0.62	0.16	1.28	0.20	0.61	-0.27	0.89

Table 10 highlights a large change of nuclear shielding constants with respect to single structure calculations ($\sigma^{opt}_{ECR,0}$, $\sigma^{opt}_{PCM,0}$) which show the largest differences compared to "$ref_{QM/MM}$". This highlights the importance of using both structural ensembles from high-level AIMD simulations as well as, at least some explicit solvation. Even the use of the "local" solvation approach over the "0" solvation approach changes cuts the difference in half for most nuclei with the PCM background solvation

benefitting more than the EC-RISM background solvation from the inclusion of a few explicit water molecules.

As the experimental observable, the chemical shift, is defined with respect to a reference compound such as DSS, the theoretical calculations require the same referencing process. This increases the computational demand but can also be beneficial as error cancellation due to insufficient level of theory might occur leading to more accurate chemical shifts compared to nuclear magnetic shielding constants.

The choice of the explicit solvent based on a distance criterion here leads to an accurate referencing where the surrounding solvent is modelled similar between an amide proton and a methyl group. This contrasts with a referencing approach where a constant number of solvent molecules is chosen. Here a size difference between solute molecules could cause errors.

An overview of the resulting chemical shifts Is shown in Figure 15 and Table 11.

Table 11: Chemical shift differences with respect to the experimental chemical shift / ppm measured by H.R. Kalbitzer[265] for the analyzed nuclei shown in Figure 15. Chemical shifts are directly referenced to DSS using the same method and averaged over magnetically equivalent atoms. For a 4 Å "full" (AIMD ensemble) explicit solvation shell augmented by an EC-RISM ("ECR"), PCM, and QM/MM (corresponding to "$ref_{QM/MM}$" in the figures) background and corresponding 2.5 Å "local" data; EC-RISM and PCM results for AIMD ensembles with 0 explicit water molecules (subscript "0") and for singular optimized structures from OLYP/6-311+G(d,p)/PCM (superscript "opt", subscript "0"). All NMR calculations used the OLYP/6-311+G(d,p) level of theory.

Nucleus	δ_{exp}	$\Delta\delta^{AIMD}_{QM/MM,full}$	$\Delta\delta^{AIMD}_{QM/MM,0}$	$\Delta\delta^{AIMD}_{ECR,full}$	$\Delta\delta^{AIMD}_{ECR,local}$	$\Delta\delta^{AIMD}_{ECR,0}$	$\Delta\delta^{opt}_{ECR,0}$	$\Delta\delta^{AIMD}_{PCM,full}$	$\Delta\delta^{AIMD}_{PCM,local}$	$\Delta\delta^{AIMD}_{PCM,0}$	$\Delta\delta^{opt}_{PCM,0}$
cis-NMA_{NH}	7.10	-0.43	-1.16	-0.39	0.07	-0.67	-1.04	-0.40	-0.52	-1.73	-2.06
trans-NMA_{NH}	7.84	-0.56	-1.51	-0.40	0.06	-0.87	-1.14	-0.38	-0.69	-2.04	-2.22
$TMAO_H$	3.25	0.10	0.05	0.09	0.14	0.05	-0.06	0.10	0.08	-0.03	-0.13
$TMAO_C$	62.18	-1.35	1.87	-1.42	0.40	2.55	2.32	-1.34	0.08	2.69	2.56

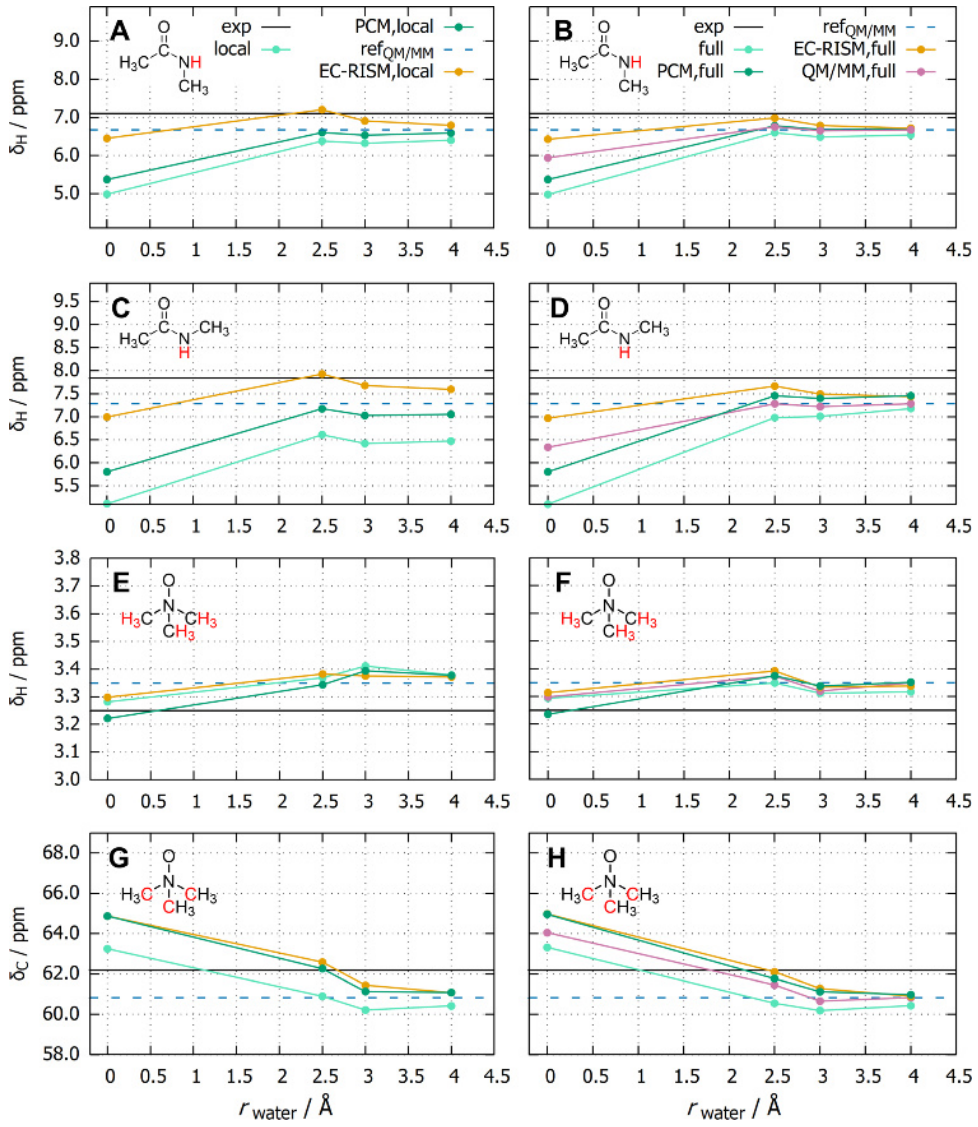

Figure 15: Chemical shifts of the highlighted nuclei (in red), averaged over magnetically equivalent atoms, with increasing radius of explicit solvent selected by local (left column) and full (right column) selection schemes. Different background models are annotated by subscripts, using same shell radii for target and reference. For ^{13}C shifts of TMAO (**G**, **H**) the distance criterion was applied to the hydrogen atoms of the methyl groups. The experimental chemical shifts are shown by a horizontal black line. "ref$_{QM/MM}$" corresponds to the "full" selection scheme at 4 Å and is also shown as dashed blue line for easier comparison.

Figure 15 and Table 11 show that depending on the used solvent model quantitative agreement with experiments can be reached. It also clearly shows the importance of using ensemble-based calculations to reach highly accurate chemical shifts as using singular optimized structures ($\Delta\delta^{opt}_{ECR,0}$) yields large errors compared to experiments. This is especially true for the amide protons where even with EC-RISM based on optimized single structures errors of more than 1 ppm occur. This error doubles to more than 2 ppm for PCM.

The overall best agreement is reached for 2.5 Å of local explicit solvation (denoted as “local” in Table 11) together with EC-RISM background solvation. Even though this is likely due to error compensation together with a relatively low level of theory used in the calculations this solvent model is a sweet spot in terms of computational demand and level of accuracy. A further increase of explicit solvent molecules leads to no significant change in the calculated chemical shifts. On the other hand, pure EC-RISM solvation (further denoted with the index “0”) is also not sufficient and the addition of on average only one explicit water molecule improves the predicted chemical shift drastically. The direct interaction between the amide proton and an explicit water molecule is crucial to prediction and needs to be accounted for explicitly. However, without any explicit solvent EC-RISM outperforms PCM especially for the amide protons by more than 1 ppm highlighting the effect that hydrogen bonds are still modelled within EC-RISM based on different local solvent site densities.

For TMAO the difference between different solvent models is way smaller. Especially for ^{1}H the range of predicted chemical shift is less than 0.2 ppm with all methods showing great agreement with experiment. This highlights the error compensation taking place in the calculation of the chemical shift as the primary information, the shielding constants, vary by more than 0.5 ppm for ^{1}H. The solvation of the methyl groups of TMAO and DSS seems to be very similar, and the error introduced by using less explicit solvation and by varying the background solvation effects both molecules in a similar way. This is also true for using single structures which in terms of chemical shifts still show good results based on an error cancellation of the shielding constants.

With the “full” solvation model it is possible to also get the chemical shift for all other nuclei of NMA. These are shown in Table 13. The additional shielding constants used for this calculation are shown in Table 12.

Table 12: Isotropic shielding constants / ppm for all analyzed nuclei, averaged over magnetically equivalent atoms, for a 4 Å "full" (AIMD ensemble) explicit solvation shell augmented by an EC-RISM ("ECR"), PCM, and QM/MM (corresponding to "$ref_{QM/MM}$" in the figures, relative to which all other numbers are reported) background and corresponding 2.5 Å "local" data; EC-RISM and PCM results for AIMD ensembles with 0 explicit water molecules (subscript "0") and for optimized structures from OLYP/6-311+G(d,p)/PCM (superscript "opt", subscript "0"). Raw data is published in the electronic supplementary material, see chapter 6.1.

Nucleus	$\sigma^{AIMD}_{QM/MM,full}$	$\Delta\sigma^{AIMD}_{QM/MM,0}$	$\Delta\sigma^{AIMD}_{ECR,full}$	$\Delta\sigma^{AIMD}_{ECR,0}$	$\Delta\sigma^{opt}_{ECR,0}$	$\Delta\sigma^{AIMD}_{PCM,full}$	$\Delta\sigma^{AIMD}_{PCM,0}$	$\Delta\sigma^{opt}_{PCM,0}$
trans-$NMA_{Hmet,1}$	29.14	0.45	-0.02	0.33	0.67	-0.01	0.51	0.74
trans-$NMA_{Hmet,2}$	28.42	0.31	-0.01	0.26	0.58	0.00	0.32	0.55
trans-$NMA_{Cmet,1}$	157.62	3.17	-0.16	2.32	4.22	-0.07	3.37	4.35
trans-$NMA_{Cmet,2}$	152.07	2.84	0.00	2.49	4.85	0.07	2.91	4.26
trans-$NMA_{Ccarbonyl}$	11.84	0.16	-0.68	-3.16	1.44	-0.57	2.03	5.32
cis-$NMA_{Hmet,1}$	29.14	0.42	-0.02	0.27	0.47	-0.01	0.45	0.64
cis-$NMA_{Hmet,2}$	28.27	0.34	-0.06	0.21	0.46	-0.04	0.33	0.58
cis-$NMA_{Cmet,1}$	160.96	3.61	-0.25	2.64	2.03	-0.17	3.76	3.06
cis-$NMA_{Cmet,2}$	149.23	2.15	-0.24	1.27	3.90	-0.14	1.94	4.12
cis-$NMA_{Ccarbonyl}$	9.68	0.62	-1.06	-3.42	-0.91	-0.72	2.20	5.44

Table 13: Chemical shift differences / ppm compared to experimental values measured by H.R. Kalbitzer[265] for the further nuclei of NMA. Chemical shifts are directly referenced to DSS using the same method and averaged over magnetically equivalent atoms. For a 4 Å "full" (AIMD ensembles) explicit solvation shell augmented by an EC-RISM ("ECR"), PCM, and QM/MM (corresponding to "$ref_{QM/MM}$" in the figures); EC-RISM and PCM results for AIMD ensembles without explicit water molecules (subscript "0") and for singular optimized structures from OLYP/6-311+G(d,p)/PCM (superscript "opt", subscript "0"). All NMR calculations used the OLYP/6-311+G(d,p) level of theory.

Nucleus	δ_{exp}	$\Delta\delta^{AIMD}_{QM/MM,full}$	$\Delta\delta^{AIMD}_{QM/MM,0}$	$\Delta\delta^{AIMD}_{ECR,full}$	$\Delta\delta^{AIMD}_{ECR,0}$	$\Delta\delta^{opt}_{ECR,0}$	$\Delta\delta^{AIMD}_{PCM,full}$	$\Delta\delta^{AIMD}_{PCM,0}$	$\Delta\delta^{opt}_{PCM,0}$
trans-$NMA_{Hmet,1}$	1.97	0.08	0.10	0.07	0.19	-0.03	0.09	0.03	-0.07
trans-$NMA_{Hmet,2}$	2.71	0.05	0.22	0.04	0.24	0.03	0.06	0.19	0.09
trans-$NMA_{Cmet,1}$	24.37	0.34	0.83	0.70	2.22	1.21	0.75	0.74	0.79
trans-$NMA_{Cmet,2}$	28.68	1.58	2.40	1.79	3.29	1.82	1.85	2.44	2.11
trans-$NMA_{Ccarbonyl}$	177.20	-6.71	-3.21	-5.79	0.63	-3.06	-5.79	-4.97	-7.23
cis-$NMA_{Hmet,1}$	2.03	0.01	0.06	0.01	0.17	0.09	0.03	0.02	-0.04
cis-$NMA_{Hmet,2}$	2.86	0.05	0.19	0.09	0.28	0.15	0.10	0.18	0.06
cis-$NMA_{Cmet,1}$	21.45	-0.08	-0.03	0.38	1.48	2.98	0.43	-0.07	1.66
cis-$NMA_{Cmet,2}$	31.93	1.17	2.68	1.62	4.10	2.36	1.65	3.00	1.85
cis-$NMA_{Ccarbonyl}$	179.90	-7.25	-4.21	-5.98	0.36	-1.26	-6.19	-5.68	-7.89

With exception of the carbonyl group all nuclei show good agreement with experimental chemical shifts. Overall mean absolute errors of 0.18 ppm for the seven ^{1}H and 2.64 ppm for the seven ^{13}C nuclei (0.90 ppm without carbonyl) can be achieved with the "$ref_{QM\,/\,MM}$" method.

Surprisingly with pure EC-RISM solvation the problematic carbonyl atom shows good results. This is likely a coincidence based on error compensation as previously shown no convergence is reached at this point. Another argument for this coincidence is that all chemical shifts are overestimated in

contrast to all other methods were. This shows the lack of convergence especially for the DSS reference for which the shielding constant decreases with more explicit water.

As all underlying shielding constants reach a converged level with respect to the treatment of the solvent and AIMD is used to generate the structures the most likely error source remaining is the level of theory, and the resulting errors can mostly be ascribed to this.

To validate this MP2 calculations were performed. Here the inclusion of explicit water was technically not possible for the whole ensembles and an extrapolation scheme was used. This scheme takes the difference between full explicit solvation and no explicit solvation from DFT calculations and applies this difference to the MP2 calculation without explicit solvation. The results are shown in Table 14.

Table 14: Shielding constants and chemical shift differences / ppm compared to experimental values measured by H.R. Kalbitzer[265] for all analyzed nuclei derived from AIMD-ensembles using the MP2/6-311+G(d,p) level of theory in the NMR calculation with the "PCM,0" solvation method and extrapolated (additional subscript "extrap.") to a hypothetical "PCM,full" solvation. The extrapolation took the difference between "PCM,full" and "PCM,0" on the OLYP level of theory in addition to MP2 "PCM,0" shielding constants. Raw data is published in the electronic supplementary material, see chapter 6.1.

Nucleus	δ_{exp}	$\Delta\delta^{AIMD}_{PCM,0,MP2}$	$\Delta\delta^{AIMD}_{PCM,MP2,full\ extrap.}$	$\sigma^{AIMD}_{PCM,0,MP2}$	$\sigma^{AIMD}_{PCM,MP2,full\ extrap.}$
DSS_H	-	-	-	31.68	31.23
DSS_C	-	-	-	198.80	195.37
cis-NMA_{NH}	7.10	-1.74	-0.41	26.33	24.55
trans-NMA_{NH}	7.84	-2.27	-0.61	26.11	24.00
$TMAO_H$	3.25	-0.15	-0.02	28.58	28.00
$TMAO_C$	62.18	5.15	1.25	131.47	131.94
trans-$NMA_{Hmet,1}$	1.97	-0.04	0.03	29.75	29.24
trans-$NMA_{Hmet,2}$	2.71	0.16	0.03	28.81	28.49
trans-$NMA_{Cmet,1}$	24.37	2.41	2.41	172.02	168.59
trans-$NMA_{Cmet,2}$	28.68	3.10	2.50	167.02	164.19
trans-$NMA_{Ccarbonyl}$	177.20	-0.15	-0.97	21.75	19.14
cis-$NMA_{Hmet,1}$	2.03	-0.02	-0.01	29.68	29.22
cis-$NMA_{Hmet,2}$	2.86	0.11	0.03	28.71	28.34
cis-$NMA_{Cmet,1}$	21.45	1.43	1.94	175.92	171.98
cis-$NMA_{Cmet,2}$	31.93	3.30	1.95	163.57	161.49
cis-$NMA_{Ccarbonyl}$	179.90	-0.24	-0.76	19.14	16.23

Especially for the previously problematic carbonyl atoms this extrapolation works well and reduces this error to less than 1 ppm. While some nuclei get worse based on this extrapolation scheme the maximum error for carbon chemical shifts is 2.5 ppm. This is smaller than any chemical shift difference between *cis*- and *trans*-NMA and shows that such a scheme to calculate chemical shifts is suitable to

determine different conformers and even populations of conformers based on recorded NMR spectra where this determination is not possible.

Overall, the first goal of this study is already fulfilled at this point. The accuracy limit for NMR calculations in aqueous solution at the selected QM level of theory can be reached by using a “full” solvation scheme with a 4 Å water shell sampled from explicit AIMD simulations of the solution together with a QM solvation model background. For the chosen model compounds this approach yields on the order of at most 5% deviation from experiment throughout, corresponding to mean absolute errors of 0.18 ppm for the seven ^{1}H and 2.64 ppm for the seven ^{13}C nuclei (0.90 ppm without carbonyl) with the “QM/MM,full” solvation model. For the carbonyl an extrapolation scheme to the MP2 level of theory also further reduced the error, highlighting the level of theory as the main remaining error source.

In terms of the second goal of reducing computational cost with still accurate treatment of solute-water interactions a simplified short-range explicit solvation scheme can only be employed if the background model retains molecular detail, which was most accurately done by using EC-RISM background solvation.

4.2.2 Comparison between AIMD and FFMD

The computational bottleneck in the previously shown workflow however, is still by far the AIMD simulation. To reduce computational cost empirical force fields are compared based on their resulting NMR observables. With the benefits of an ensemble calculation already shown, sampling based on force fields could be an interesting alternative to demanding AIMD simulations. A comparison of shielding constants (Table 15) and chemical shifts (Table 16 and Table 17) is shown. As the “QM/MM,full” approach was not performed for all force field calculations, the comparison will focus on the EC-RISM and PCM background solvation models. Especially EC-RISM background solvation, as shown earlier, even yields faster convergence and no significant difference with the “full” explicit solvation compared to QM/MM.

Table 15: Comparison of shielding constants / ppm including statistical uncertainties for highlighted nuclei derived from AIMD- and FFMD-ensembles with "local" and "full" explicit solvation augmented by an EC-RISM- or PCM-background.

Nucleus	$\sigma^{AIMD}_{ECR,local}$	$\sigma^{FFMD}_{ECR,local}$	$\sigma^{AIMD}_{PCM,local}$	$\sigma^{FFMD}_{PCM,local}$	$\sigma^{AIMD}_{ECR,full}$	$\sigma^{FFMD}_{ECR,full}$	$\sigma^{AIMD}_{PCM,full}$	$\sigma^{FFMD}_{PCM,full}$
DSS_H	31.48 ± 0.01	31.93 ± 0.01	31.50 ± 0.01	31.95 ± 0.01	31.16 ± 0.01	31.60 ± 0.01	31.19 ± 0.01	31.63 ± 0.01
DSS_C	184.71 ± 0.13	187.96 ± 0.12	184.37 ± 0.13	187.50 ± 0.13	182.54 ± 0.13	185.65 ± 0.13	182.67 ± 0.13	185.75 ± 0.12
trans-NMA_{NH}	23.58 ± 0.07	24.94 ± 0.06	24.33 ± 0.07	25.72 ± 0.06	23.73 ± 0.07	25.00 ± 0.06	23.73 ± 0.07	25.05 ± 0.01
cis-NMA_{NH}	24.31 ± 0.07	25.25 ± 0.06	24.90 ± 0.07	25.92 ± 0.06	24.45 ± 0.08	25.58 ± 0.07	24.49 ± 0.07	25.66 ± 0.06
$TMAO_H$	28.10 ± 0.01	28.58 ± 0.02	28.16 ± 0.01	28.66 ± 0.02	27.83 ± 0.02	28.34 ± 0.02	27.84 ± 0.01	28.38 ± 0.02
$TMAO_C$	122.12 ± 0.17	127.05 ± 0.25	122.12 ± 0.17	127.04 ± 0.22	121.64 ± 0.16	127.04 ± 0.24	121.70 ± 0.16	126.80 ± 0.53

Table 16: Chemical shift differences / ppm compared to experimental values for the highlighted nuclei derived from AIMD- and FFMD-ensembles with "local" and "full" explicit solvation augmented by an EC-RISM- or PCM-background. The underlying shielding constants used to compute these values are shown in Table 15.

Nucleus	δ_{exp}	$\Delta\delta^{AIMD}_{QM/MM,full}$	$\Delta\delta^{AIMD}_{ECR,full}$	$\Delta\delta^{FFMD}_{ECR,full}$	$\Delta\delta^{AIMD}_{ECR,local}$	$\Delta\delta^{FFMD}_{ECR,local}$	$\Delta\delta^{AIMD}_{PCM,full}$	$\Delta\delta^{FFMD}_{PCM.full}$	$\Delta\delta^{AIMD}_{PCM,local}$	$\Delta\delta^{FFMD}_{PCM,local}$.
cis-NMA_{NH}	7.10	-0.43	-0.39	-1.08	0.07	-0.42	-0.40	-1.13	-0.52	-1.09
trans-NMA_{NH}	7.84	-0.56	-0.40	-1.24	0.06	-0.85	-0.38	-1.25	-0.69	-1.63
$TMAO_H$	3.25	0.10	0.09	0.01	0.14	0.09	0.10	0.00	0.08	0.02
$TMAO_C$	62.18	-1.35	-1.42	-3.71	0.40	-1.41	-1.34	-3.37	0.08	-1.86

Table 17: Chemical shift differences / ppm based on AIMD and FFMD ensembles compared to experimental values for the further nuclei of NMA.

Nucleus	δ_{exp}	$\Delta\delta^{AIMD}_{QM/MM,full}$	$\Delta\delta^{AIMD}_{ECR,full}$	$\Delta\delta^{FFMD}_{ECR,full}$	$\Delta\delta^{AIMD}_{PCM,full}$	$\Delta\delta^{FFMD}_{PCM.full}$
trans-$NMA_{Hmet,1}$	1.97	0.08	0.07	0.08	0.09	0.08
trans-$NMA_{Hmet,2}$	2.71	0.05	0.04	0.26	0.06	0.26
trans-$NMA_{Cmet,1}$	24.37	0.34	0.70	2.33	0.75	2.24
trans-$NMA_{Cmet,2}$	28.68	1.58	1.79	2.66	1.85	2.71
trans-$NMA_{Ccarbonyl}$	177.20	-6.71	-5.79	-9.64	-5.79	-9.94
cis-$NMA_{Hmet,1}$	2.03	0.01	0.01	0.04	0.03	0.05
cis-$NMA_{Hmet,2}$	2.86	0.05	0.09	0.23	0.10	0.24
cis-$NMA_{Cmet,1}$	21.45	-0.08	0.38	2.84	0.43	2.76
cis-$NMA_{Cmet,2}$	31.93	1.17	1.62	2.45	1.65	2.48
cis-$NMA_{Ccarbonyl}$	179.90	-7.25	-5.98	-10.82	-6.19	-11.31

First, it is noteworthy that the uncertainties from FFMD ensembles, as shown in Table 15, are comparable to AIMD ensembles with one minor exception for ^{13}C of TMAO, for which errors are a bit larger but still in the range of 0.25 ppm.

Table 16 shows that while for TMAO the performance of the FFMD is still as good as the AIMD the amide protons show worse results and the benefit of this sampling approach compared to single

structure calculations is basically gone (erros of -1.08, -1.24, 0.01 and -3.71 ppm for $\Delta\delta^{FFMD}_{ECR,full}$, compared to -1.04, -1.14,-0.06 and 2.32 ppm for $\Delta\delta^{opt}_{ECR,0}$, see Table 11) .

This shows that especially the force field of NMA needs a further look to determine the underlying reasons for the worse predictions. Using the "ECR,full" solvation method the amide proton shielding constants increase by 1.27 and 1.13 ppm for *trans* and *cis* respectively going from AIMD to FFMD ensembles (see Table 15). This is the main reason for the worse chemical shift prediction, as the ^{1}H DSS shielding constant only increases by 0.44 AIMD to FFMD ensembles (see Table 15).

Surprisingly the shielding constants of the ^{13}C DSS and TMAO also show large differences between AIMD and FFMD ensembles of more than 3 ppm, however here error cancellation plays a huge role and leads to still accurate chemical shifts shown in Table 16 and even for the further ^{13}C of NMA shown in Table 17 FFMD errors compared to experiment as well as AIMD are less than 3 ppm with the exception of the carbonyl where the level of theory, as shown before in Table 14, leads to larger errors of now more than 10 ppm based on FFMD ensembles.

4.2.3 Structural comparison between AIMD and FFMD

To determine the reason why FFMD based predictions are worse than AIMD predictions structural features of the solvent and the solute are compared for both methods. In Figure 16 the radial distribution functions (RDFs) are shown surrounding all nuclei shown in Figure 13.

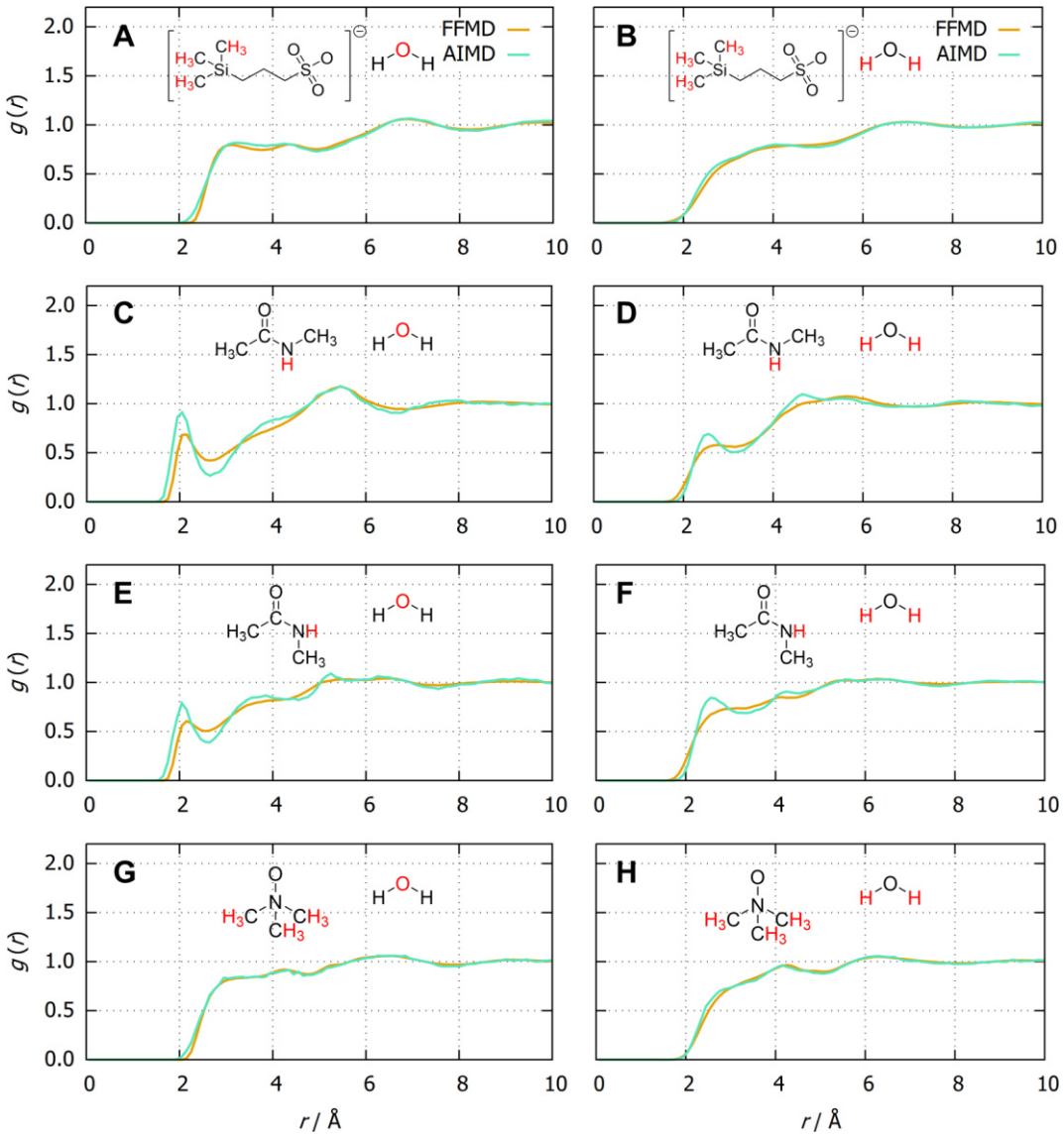

Figure 16: Radial distribution functions of the highlighted water nuclei surrounding the highlighted solute nuclei for which shielding constants have been calculated. RDFs from FFMD (orange) and AIMD (blue) are compared.

The RDFs around DSS and TMAO are similar for the AIMD and FFMD. This explains why for TMAO even the FFMD can be used to get good NMR predictions. The RDFs are similar for both the near range showing that the solute parameters are well chosen as well as the long range which shows that the water force field also resembles the AIMD to a high degree.

In contrast a clear difference can be seen for the amide proton of NMA. The hydrogen bond which is crucial for accurate chemical shift calculations is modelled weaker for the FFMD. The first peak of the oxygen RDF (**C,E**) is ca. 0.1 Å closer to the proton for AIMD and the density is also higher. Based on the isolated model calculations shown in Figure 7 this fact alone should reduce the shielding by ca. 0.6 ppm. This is in line with the data in Table 15 where the difference with the "local" EC-RISM solvation is 1.36 and 0.99 ppm respectively for *trans*- and *cis*-NMA but not fully explains the difference. To

further explain the larger difference in predicted chemical shifts intramolecular parameters in the H-N-C-O dihedral angle (**A**) and the N-H bond length (**B**) are shown in Figure 17. A correlation analysis for between both structural parameters and the resulting NMR parameters follows in Figure 18.

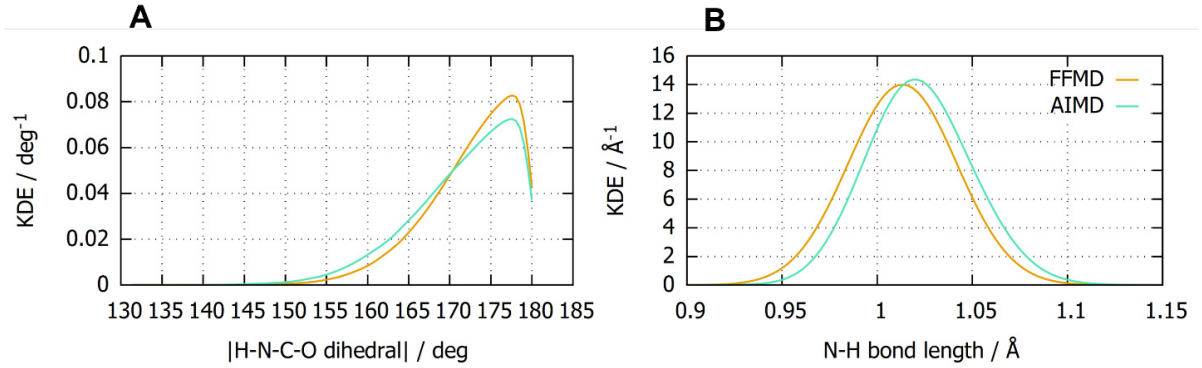

Figure 17: Kernel density estimate (KDE) of the absolute H-N-C-O dihedral angle (**A**) and the N-H bond length (**B**) for NMA in the *trans* conformation simulated by FFMD (orange) and AIMD (blue). For the analysis, 300000 snapshots of either simulation were considered. For the KDE a bandwidth of 1 deg and 0.01 Å was used, yielding average values of 171.52 deg and 1.022 Å for AIMD and 172.64 deg and 1.013 Å for FFMD.

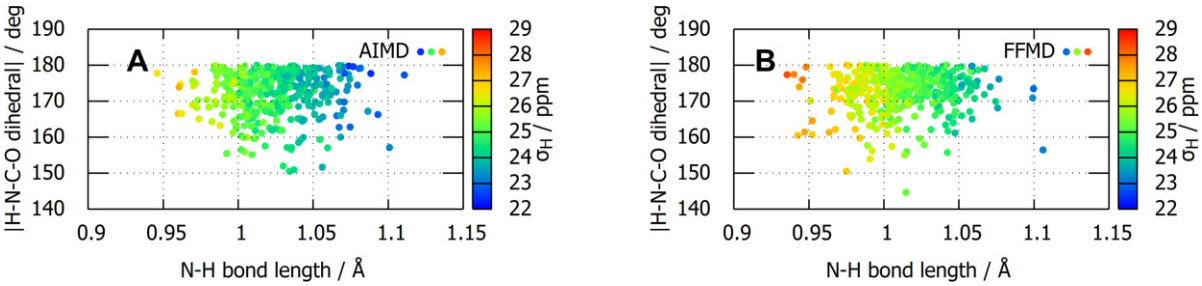

Figure 18: Heatmap of the amide shielding constants of *trans*-NMA in dependence of the absolute H-N-C-O dihedral angle and the N-H bond length for the snapshots from AIMD (**A**) and FFMD (**B**) used in the NMR calculations. Here the "EC-RISM,0" solvation method was employed, yielding average shielding constants of 24.65 ppm (AIMD) and 25.41 ppm (FFMD).

Figure 17 shows that there is a measurable difference between the intramolecular structural parameters for NMA based on AIMD and FFMD ensembles. The absolute dihedral angle for the AIMD is ca. 1 ° smaller showing a higher flexibility of in the AIMD and the N-H bond length which was earlier (Figure 7) shown to directly correlate with NMR parameters is ca. 1 pm larger for the AIMD. A correlation analysis in Figure 18 shows a clear dependence of the shielding constants on the N-H bond length where higher values of more than 1.05 Å are much more common in the AIMD. Here the difference of 0.76 ppm almost perfectly explains the remaining difference between chemical shifts from AIMD and FFMD even at the fully solvated system. This shows that for the difference between AIMD and FFMD roughly 60 % can be attributed to the difference in the molecular structures and the remaining 40 % are due to the difference in explicit solvation. So, a combination of these two differences in the generated structures fully explains the difference between AIMD and FFMD. With

the resulting chemical shifts from AIMD being closer to experimental ones the structural parameters from the AIMD ensembles are more trustworthy and likely closer to a realistic depiction of NMA. With only two structural parameters identified as responsible for the error in the NMR prediction of the amide proton more methods can be tested and evaluated based on the solute and solvent structures they produce. The NMR parameters using a similar approach hybrid solvation approach should then follow with predictable quality. Remaining errors based on the level of theory could still remain.

4.2.4 Simulation of NMA with Machine Learning Potentials

As another method of generating structural ensembles MLPs (MLP) in form of the ANI-type potentials were investigated. They should potentially yield more accurate results compared to FFMD ensembles at a computational cost that is ca. 50 times higher but still three to four orders of magnitude less compared to the AIMD. The models that are investigated are detailed in Table 1. This investigation started initially on the ANI-2x[66] and ANI-dr[237] model however, another model here called ANI-custom was also investigated, as it was trained with additional datapoints of alanine-dipeptide water clusters, which should translate well to the study of NMA, but is in contrast to the general purpose approach of the ANI models.

For the ANI-type potentials neutral molecules containing the elements H, C, N, O, S, F and Cl were used in the training data. Thus, molecules like DSS containing Silicon cannot be simulated with this potential and even molecules like TMAO with a high charge density are unlikely to give decent results as no N-oxide molecules were present in the training data. The focus of this proof of concept is therefore on the different conformers of NMA for which AIMD and FFMD data was generated and is used as a comparison.

The simulation setup for these was kept the same as for the AIMD and FFMD with one NMA molecules surrounded by 152 water molecules. The density was determined individually, as sometimes large deviations could occur as shown in Table 18.

Table 18: Densities used in the NVT simulation of NMA. The density from FFMD was ales used for AIMD simulations. It is noteworthy that the ANI-based densities are based on 100 ps *NpT* simulations while the FFMD used 1.5 ns *NpT* simulations. All simulations used one solute molecule together with 152 water molecules.

solute	ρ_{FFMD} / kg/m^3	$\rho_{ANI\text{-}2x}$ / kg/m^3	$\rho_{ANI\text{-}dr}$/ kg/m^3	$\rho_{ANI\text{-}custom}$ / kg/m^3
trans-NMA	998.074	970.19	875.45	874.27
cis-NMA	998.074	959.49	894.28	876.50

A comparison of the resulting structural parameters of ANI-2x and ANI-dr is shown in Figure 19. The improved solvent distributions from ANI-custom around *trans*-NMA are shown in Figure 20.

To benchmark the resulting NMR data, nuclear magnetic shielding constants are given in Table 19.

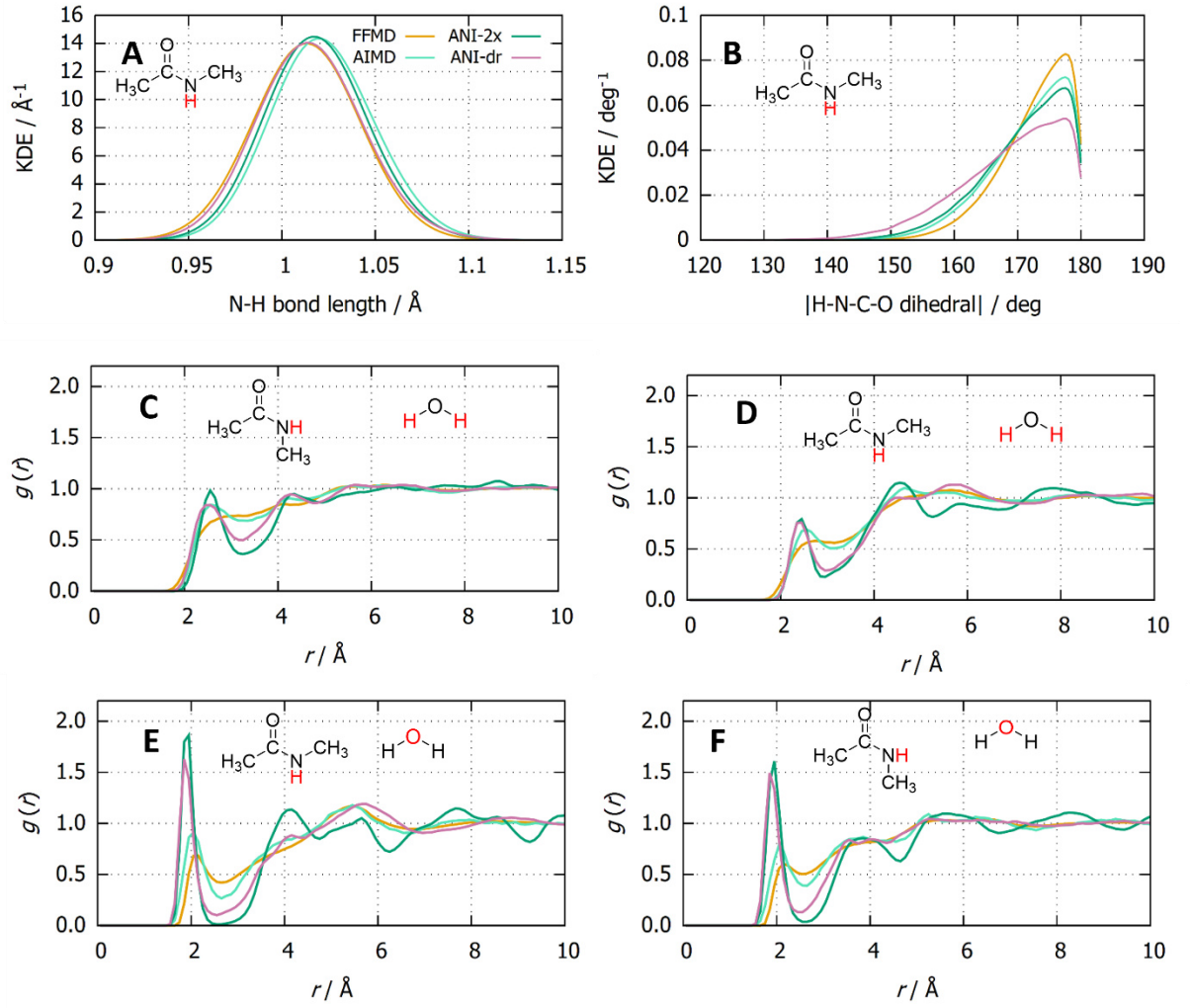

Figure 19: Intramolecular parameters of *trans*-NMA (**A**, **B**) and intermolecular RDFs for *trans*- and *cis*-NMA from AIMD, FFMD and MLMD using the ANI-2x[66] model and another model trained with dispersion[107] and repulsion[236] (ANI-dr).[237]

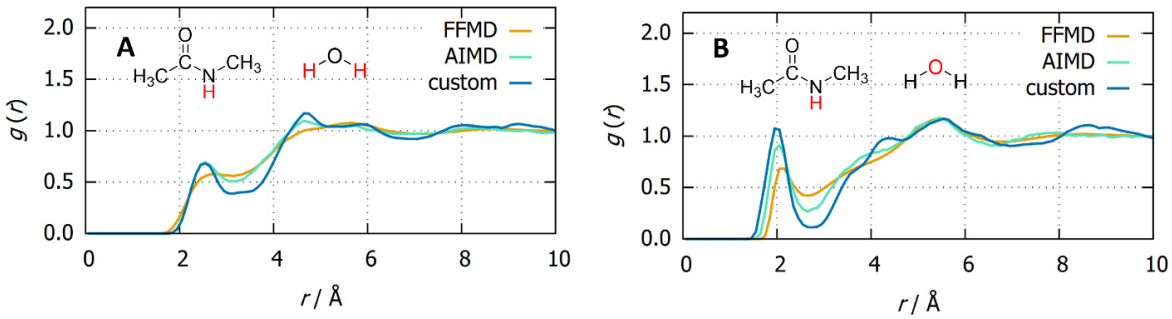

Figure 20: Radial distribution functions of the highlighted water nuclei surrounding the highlighted solute nuclei for which shielding constants have been calculated. RDFs from FFMD (orange) and AIMD (blue) are compared with an MLMD using a custom MLP provided by J. Xue[239] with additional training data of alanine dipeptide-water clusters.

Table 19: Isotropic shielding constants / ppm for all nuclei of *trans*- and *cis* NMA, averaged over magnetically equivalent atoms with respect to AIMD simulations for all MLMD simulations. All models used the "$ref_{QM/MM}$" solvation method. FFMD values are also given as a reference. The MAEs with respect to the AIMD simulation are given for all hydrogen and carbon atoms respectively. Raw data is published in the electronic supplementary material, see chapter 6.1.

Nucleus	$\sigma^{AIMD}_{ref,QM/MM}$	$\Delta\sigma^{ANI\text{-}dr}_{ref,QM/MM}$	$\Delta\sigma^{ANI\text{-}2x}_{ref,QM/MM}$	$\Delta\sigma^{ANI\text{-}custom}_{ref,QM/MM}$	$\Delta\sigma^{FFMD}_{ref,QM/MM}$
trans-NMA_{Hamide}	23.90	0.12	-0.11	0.19	1.34
trans-$NMA_{Hmet,1}$	29.14	0.55	0.02	0.33	0.44
trans-$NMA_{Hmet,2}$	28.42	0.21	-0.34	0.15	0.24
trans-$NMA_{Cmet,1}$	157.62	3.15	2.16	2.50	1.53
trans-$NMA_{Cmet,2}$	152.07	3.40	1.63	3.25	2.27
trans-$NMA_{Ccarbonyl}$	11.84	3.82	6.02	3.33	7.08
cis-NMA_{Hamide}	24.51	0.05	1.04	-0.42	1.25
cis-$NMA_{Hmet,1}$	29.14	0.47	-0.70	0.36	0.41
cis-$NMA_{Hmet,2}$	28.27	0.38	-0.02	0.21	0.29
cis-$NMA_{Cmet,1}$	160.96	2.66	-0.89	2.85	0.68
cis-$NMA_{Cmet,2}$	149.23	3.60	1.50	2.82	2.13
cis-$NMA_{Ccarbonyl}$	9.68	2.49	0.08	0.52	8.08
MAE_H	-	0.30	0.37	0.28	0.66
MAE_C	-	3.19	2.04	2.54	3.63

Figure 19 shows that especially the intramolecular parameters in the N-H bond length and the amide dihedral angle are remarkably well represented using the ANI-2x model and very close to the AIMD. However, it shows a worse representation of the explicit solvent. Here a water molecule is basically stuck to the amide proton at a low distance and diffusion away is not present. To illustrate this the average standard deviation of the NH-O_{water} bond length was measured for a particular water molecule over the simulation. For ANI-2x a value of 0.29 Å was calculated, showing that this water molecule is only present within the first peak shown in the RDF (Figure 19, **E**). This is even true over a simulation length five times longer than the AIMD simulation time where diffusion occurred. For ANI-custom and ANI-dr this value increases to 0.64 and 1.08 Å showing improved diffusion and explaining the improved

RDFs. Even for longer distances the ANI-2x RDFs show distinct peaks and valleys which correspond to a single fixed placement of water molecules which does not change during the simulation. For the ANI-dr model the description of the solvent improves due to the improved diffusion. However, the description of the intramolecular parameters is worse and more closely resembles the FFMD due to the level of theory which was used, B97-3c yielding worse results. As expected, the ANI-custom more closely resembles the RDFs from the AIMD. Especially the H-H RDF shows great agreement. For both RDFs the first peak now closely resembles the height of the AIMD but for the O-H RDF it is still a bit closer compared to the AIMD, similar to what was found for the other ANI-models. With smaller valleys and in general a smoother RDF the custom model shows the signs of improved diffusion during the simulation.

For the resulting shielding constants shown in Table 19, the custom model shows the best agreement with AIMD for the ^{1}H, due to a well-balanced prediction of both the methyl and amide protons. This shows that the addition of few, customized datapoints is a promising way to specifically improve on a general-purpose model. Overall, the use of the custom MLP halves the error of the ^{1}H shielding constants of the FFMD with respect to the AIMD-based ensembles. For ^{13}C ANI-2x shows the best agreement with an MAE of 2.04 ppm, yielding overall the best agreement with AIMD simulations. Here the improvement compared to the FFMD is not as large but with 1.5 ppm still significant. Overall, all MLPs showed an improvement compared to the FFMD for all nuclei.

With one of the use cases of NMR predictions being the identification of different conformers, all methods are also evaluated in this regard by comparing chemical shift differences between *trans*- and *cis*-NMA with experiments. These are shown in Table 20.

Table 20: Chemical shift differences / ppm between *trans*- and *cis*-NMA using ensembles from AIMD, FFMD and MLMD using the ANI-2x model and another model trained with dispersion[107] and repulsion[236] (ANI-dr).[237] For all calculations the "ref$_{QM/MM}$" solvation method was used including 4 Å of explicit solvent around the whole molecule and additional TIP3P-point charges[241] were placed on the water molecules outside of the explicit QM zone.

Nucleus	$\Delta\delta_{exp}$	$\Delta\delta^{AIMD}$	$\Delta\delta^{ANI\text{-}dr}$	$\Delta\delta^{ANI\text{-}2x}$	$\Delta\delta^{ANI\text{-}custom}$	$\Delta\delta^{FFMD}$
NMA_{amide}	0.740	0.611	0.371	1.200	-0.004	0.522
$NMA_{Ccarbonyl}$	-2.700	-2.155	-5.624	0.084	-4.976	-1.163
$NMA_{Hmet,1}$	-0.060	0.007	-0.113	0.061	0.036	-0.027
$NMA_{Hmet,2}$	-0.150	-0.148	-0.230	-0.296	-0.083	-0.101
$NMA_{Cmet,1}$	2.920	3.339	1.734	2.621	3.690	2.484
$NMA_{Cmet,2}$	-3.250	-2.840	-3.798	-3.998	-3.274	-2.985

In the differentiation between conformers ANI-dr is noteworthy as it yields the right trend for all nuclei which would allow a perfect assignment of peaks to the corresponding conformers. This is also true

for the FFMD but even the AIMD misassigns one peak ($H_{met,1}$). However the absolute differences are best match by the AIMD, as expected, with the FFMD yielding the second best differences.

4.2.5 MLMD-refinement of FFMD ensembles

As outlined before, the MLMD simulations are a promising tool, yielding overall more accurate absolute shielding constants compared to the FFMD (Table 19), but they are still showing significant flaws especially with regards to the solvent distributions (Figure 19) and calculated densities (Table 18) and suffer in terms of chemical shift differences between conformers (Table 20). Here even cheap FFMD simulations yielded better agreements with AIMD simulations. However, especially ANI-2x showed great agreement in the intramolecular description (Figure 19 **A, B**). In an attempt to combine the benefits of both the FFMD and MLMD simulations the initial FFMD ensemble, represented with 400 snapshots, with an accurate solvent distribution and density was taken. Then very short MLMD simulations of up to 200 fs were performed using the ANI-2x model. The goal was to mostly improve intramolecular parameters while still, mostly, maintaining the right solvent distribution.

Based on the previous findings that mainly showed an improved of intramolecular parameters (Figure 19) an interesting concept is to start with an initial ensemble from the FFMD and only use short simulations with the MLMD to improve only the intramolecular parameters such as bond lengths but not suffer from the lack of diffusion within the MLMD, that was especially visible using the ANI-2x model. Here an ensemble of 400 starting geometries was used and each refined for up to 200 fs using the ANI-2x model. At different simulation times of each structure snapshots were taken and used to calculate nuclear magnetic shielding constants.

The resulting average nuclear magnetic shielding constants of *trans*-NMA with increasing time are shown in Figure 21 and listed in Table 21.

Table 21: Isotropic shielding constants / ppm for all nuclei of *trans*- NMA, averaged over magnetically equivalent atoms with respect to AIMD simulations for an increasing time of ANI-2x refinement. The simulation started with the FFMD ensemble. The "$ref_{QM/MM}$" solvation method was used throughout all calculations. FFMD values are also given as a reference. The MAEs with respect to the AIMD simulation are given for all hydrogen and carbon atoms.

Nucleus	$\sigma^{AIMD}_{ref,QM/MM}$	19 fs	39 fs	119 fs	199 fs	$\Delta\sigma^{FFMD}_{ref,QM/MM}$
trans-NMA_{Hamide}	23.90	0.94	0.84	0.90	0.87	1.34
trans-$NMA_{Hmet,1}$	29.14	0.31	0.13	0.03	0.09	0.44
trans-$NMA_{Hmet,2}$	28.42	0.07	-0.02	0.03	-0.13	0.24
trans-$NMA_{Cmet,1}$	157.62	-0.20	1.24	-0.31	0.69	1.53
trans-$NMA_{Cmet,2}$	152.07	0.84	1.23	2.43	1.54	2.27
trans-$NMA_{Ccarbonyl}$	11.84	3.20	4.12	4.05	3.50	7.08
MAE_H	-	0.44	0.33	0.32	0.37	0.67
MAE_C	-	1.41	2.20	2.26	1.91	3.63

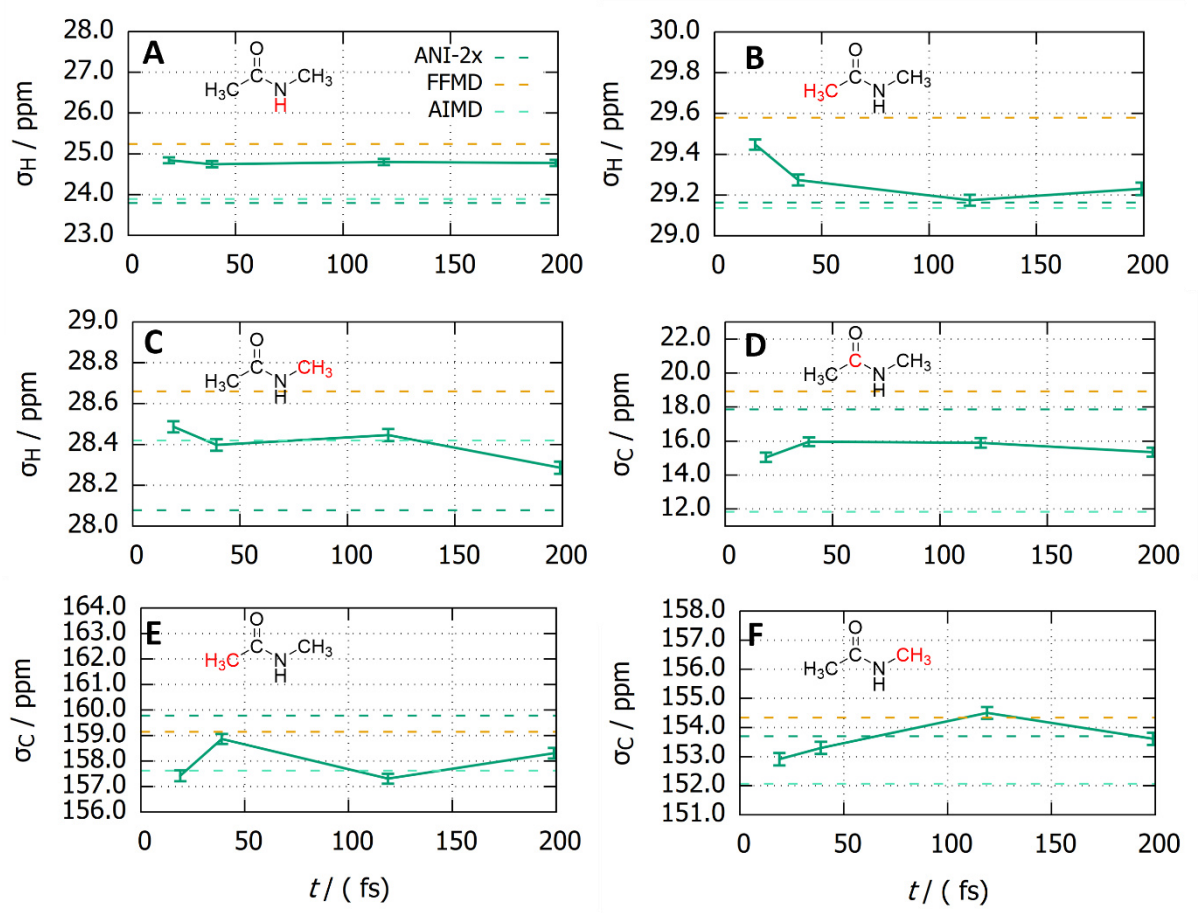

Figure 21: Time dependent shielding constants for the highlighted nuclei (red) of *trans*-NMA. Starting from the whole FFMD-ensemble every snapshot was simulated using the ANI-2x MLP.[66] Snapshots were taken from each trajectory at the given time. Average values from the FFMD, AIMD and ANI-2x MLP itself are shown for reference as dashed lines. For all calculations the "ref$_{QM/MM}$" solvation method was used which includes 4 Å of explicit solvent around the whole molecule and additional TIP3P-point charges[241] were placed on the water molecules outside of the explicit QM zone. Raw data is published in the electronic supplementary material, see chapter 6.1.

Figure 21 shows that all nuclear shielding constants get closer to the AIMD values using the ANI-2x refinement. This validates this proposed strategy as a computationally cheap way to improve FFMD based NMR predictions. This strategy uses only at most 80 ps of total MLMD simulation time and improves ^{1}H shielding predictions by ca. 0.3 ppm and ^{13}C predictions by ca. 1.5 ppm on average. This improvement can already be found after 19 fs (see Table 21) which only requires 8 ps of MLMD simulation time, however oscillations in the calculated shielding constants are still visible, making this prediction unreliable without further investigations. An alternative way that was not investigated here is to optimize the FFMD snapshots using ANI-2x, this would further reduce the computational demand but could again introduce some of the flaws from the MLP that were shown before.

The largest change due to the MLMD refinement is found for the problematic carbonyl shielding constant which improves by 5 ppm even outperforming the pure ANI-2x simulation. Especially for the

amide proton a larger error of almost 1 ppm remains. This is due to the worse modelling of the explicit solvent around the amide proton already shown in Figure 19. With this refinement method it cannot be expected to improve the explicit solvation and therefore errors based on this will remain. Overall, the NMR predictions compared to the full ANI-2x simulation shown in Table 19 improve.

To show how this refinement acts on the intramolecular parameters Figure 22 show the change in the average bond length with increasing MLMD simulation time.

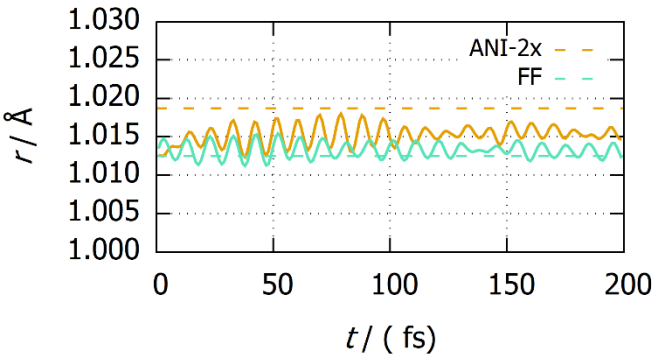

Figure 22: Time dependent N-H bond length for the N-H bond of *trans*-NMA. Starting from the whole FF-ensemble every snapshot was simulated using the ANI-2x MLP. Snapshots were taken from each trajectory at the given time and the bond length was averaged over 400 snapshots. Average values from the whole FFMD (blue) and the previously performed full 1 ns MLMD(orange) using ANI-2x are shown for reference as dashed lines.

Figure 22 shows that the average bond length increases when the ANI-2x MLP is used to run the MD simulation. Starting from the FFMD value an increase of ca. 0.5 pm can be seen after 200 fs. This also seems to be a converged value as only periodic changes around a value of ca. 1.016 Å remain which are similar for both the MLMD and FFMD. These are due to a nonuniform velocity distribution for the 400 snapshots used at the start of this simulation and propagate over this small simulation time.

A further increase in the N-H bond length to get closer to the equilibrated value for the MLMD is only expected for longer time scales where the explicit solvent rearranges. However, in this approach this was non-desirable to keep up the MLMD simulation time short. At a given point running the full 1 ns ANI-2x simulation that was performed in chapter 4.2.4 becomes the more reasonable option. However, with an N-H bond length of 1.016 Å the difference between FFMD and AIMD could already be decreased by 30 % by just running 80 ps of MLMD simulations in total.

4.2.6 Summary and Outlook

This study was set up with two goals, which could be reached. The first was to reach the maximum accuracy with respect to the solvent description based on a combination of AIMD simulations from

which snapshots were taken and NMR parameters were calculated using explicit solvation, together with EC-RISM, PCM or QM/MM background solvation. Models systems NMA, in *cis*- and *trans*-conformation and TMAO were used to investigate both polar and apolar solute-solvent interaction while the experimental reference molecule DSS was used to fully replicate the experimental chemical shift determination.

As a first insight a convergence with respect to the number of structures taken from MD simulations was found for roughly 200 snapshots, while statistical uncertainties of 0.01 ppm for ^{1}H and 0.1 ppm for ^{13}C chemical shifts can be reached with roughly 400 snapshots, which could be reduced if multiple equivalent nuclei are present in the molecule.

Overall, a convergence with respect to the size of the explicit solvation shell was found at a distance of 4 Å around the solute, determining this as an accurate ("full") solvent description. For the chosen model compounds this approach yields on the order of at most 5% deviation from experiment throughout, corresponding to mean absolute errors of 0.18 ppm for the seven ^{1}H and 2.64 ppm for the seven ^{13}C nuclei (0.90 ppm without carbonyl) with the "QM/MM,full" solvation model. With this benchmark level achieved further, computationally less demanding methods could be clearly characterized.

As chemical shifts are measured for specific nuclei, an even more restrictive local solvation approach was investigated which includes solvent molecules only up to a distance of 2.5 Å around the nucleus of interest. Here discrepancies were found especially for the amide protons of NMA which suffered the most from incomplete solvation even around the opposing carbonyl group.

However, this error could drastically be reduced most efficiently by the inclusion of an EC-RISM background solvation. This "local,ECR" solvation approach on average only took ca. 1 explicit water molecule (compared to almost 30 for the "full" explicit solvation approach) but yielded very similar results for the chemical shifts at a much cheaper computational cost. The 1 explicit water molecule, however, was deemed to be very necessary to be included highlighting the importance of close interactions which, in this magnitude could not be replicated by pure EC-RISM solvation.

Using different background solvation methods such as PCM yielded significantly larger chemical shift errors, while a QM/MM approach converges to the same values as the "local" EC-RISM solvation but significantly slower reducing the benefit of the "local" solvation.

For further investigations the "local" explicit solvation together with EC-RISM background solvation is especially interesting for larger molecules such as peptides or even proteins where specific nuclei are of interest. Here taking only the local explicit solvent molecules drastically reduces the computational cost of respective NMR calculations at, based on the model system NMA, only marginally increased errors.

In an attempt to further reduce the computational cost, the use of FFMD simulations compared to AIMD simulations was found to introduce significant errors, mostly for the amide proton which could be explained by a too short N-H bond length combined with a too weak hydrogen bond compared

with AIMD simulations. Both factors yielding a lower chemical shift for FFMD simulations compared to AIMD simulations.

For other nuclei the FFMD simulations yielded comparable results with AIMD simulations which could be explained by significant error compensation by directly referencing to DSS, which was significantly impacted by the force field description. In this way it is also possible to further benchmark and improve force fields on easily and accurately measurable NMR parameters benchmarked against calculated NMR parameters derived from their structural ensembles using the most accurate solvent description found here.

As a third option MLMD simulations based on a variety of ANI-type MLPs were investigated for NMA. Depending on the specific potential used they could closely match AIMD-based structural ensembles. The ANI-2x model could closely match the N-H bond length from AIMD, however it lacked an accurate description of the explicit solvent around NMA due to too little diffusion of solvent. A custom MLP introducing alanine dipeptide water clusters to the training data could improve the description of solvent close to the solute which also lead to better agreement of the ^{1}H-NMR parameters.

Overall, MLMD simulations are a promising tool, yielding a higher accuracy on calculated shielding constants with errors of ca. 0.3 and 2.5 ppm for ^{1}H and ^{13}C with respect to AIMD-based shielding constants, cutting the FFMD errors in half. But they also have a higher computational cost compared to FFMD simulations and limitations based on the training data still remain.

As a last method an MLMD refinement of the FFMD ensembles was investigated for *trans*-NMA. Here, starting from a set of 400 snapshots within femtoseconds intramolecular parameters such as bond lengths could be relaxed close to the MLMD equilibrium while the solvent surrounding the solute still retains the fairly accurate FFMD structure. An improvement in NMR parameters for all nuclei of NMA was found with all of them more closely matching the AIMD data compared to the initial FFMD data.

The computational cost of running short femtosecond MLMD simulations even for hundreds of snapshots was still low, making this a viable option to improve all structural ensembles from FFMD.

To sum up this study, it was shown that a high level of accuracy in the NMR predictions can be reached using significant computational effort for both AIMD simulations, and ensemble-based NMR calculations using explicit solvation. All computationally less demanding methods investigated also yielded less accurate NMR predictions, however this study yielded a broad overview on the application of a multitude of methods and what could be expected from them and identified sweet spots in both accuracy and cost. This allows for an informed decision in all further studies based on the complexity of the investigated system, the computational resources that are available and the necessary accuracy to answer the specific research question.

4.3 High-pressure NMR measurements

Even though the main focus of this thesis is based on theoretical investigations of magnetic resonance parameters, NMR experiments using high hydrostatic pressure have also been performed. The main use of these lies in the benchmarking of theoretical methods even at extreme conditions such as high hydrostatic pressure where experimental data is limited. Therefore, it was of interest to first quantify the uncertainty of pressure-dependent NMR measurements. As the perturbance due to high hydrostatic pressure is small any benchmark measurement should be reliable, which could be challenging. To investigate the uncertainty, repeated measurements of alanine have been made among multiple spectrometers and with or without the addition of Tris-HCl buffer. The pressure-dependent ^{1}H chemical shifts of the methyl group and the alpha-H of alanine are shown in Figure 23.

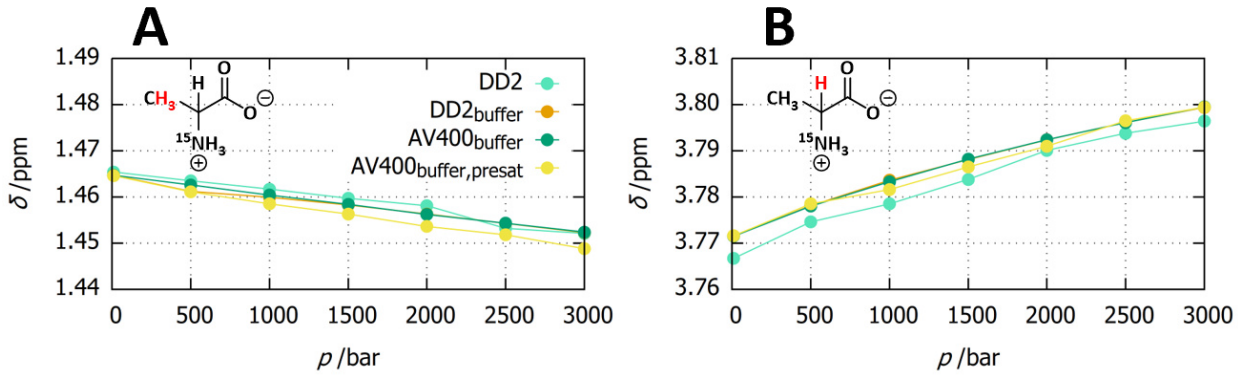

Figure 23: Pressure-dependent ^{1}H chemical shift of the methyl group (A) and the α-H (B) of alanine using two different spectrometers (DD2 and AV400). For both methods a tris-HCl buffer solution was added and using the AV400 probe the addition of solvent suppression was investigated (presat). For all results an internal reference of DSS at 0 ppm at all pressures was used.

Figure 23 shows a consistent trend of a decreasing chemical shift for the methyl group and increasing chemical shift of the alpha-H. Here the difference between individual measurements using two different spectrometers (AV400 and DD2), with and without buffer and using two different pulse sequences including water suppression (presat) is marginal and still significantly smaller than the change in chemical shift induced by pressure. The largest factor still is the addition of buffer as this changes the constitution of the specific sample, however even this effect is only marginal. This highlights the fact that pressure-dependent NMR measurements are a reliable source of benchmark data for theoretical calculations and will be used as such in the following chapters of this thesis.

Additional factors come into play for the ^{13}C-and ^{15}N-measurements in terms of the method of referencing chemical shift data. In terms of using the data as benchmark data for theoretical calculations especially for ^{13}C, using the methyl groups of DSS as a pressure-dependent internal standard in the same way as for ^{1}H yields the closest methodological correspondence between theory

and experiment. Taking the Ξ value corresponding to the resonance frequency of the ^{1}H-signal of DSS can only be done for the pressure-dependent part of the calculated chemical shift as done in Ref.[8] However part of this experimental study was to further investigate different referencing methods especially for ^{13}C, for which internal DSS is still present.

Here the reference frequency for the chemical shift from equation (81) can be defined in multiple ways. Previously an indirect reference frequency defined by

$$\nu_0^{13C}(p) = \Xi(1\ \text{bar})\nu_0^{1H,DSS}(p) \tag{101}$$

was used and was even used in comparison to directly referenced, calculated chemical shifts.[7] This is only comparable under the assumption that the pressure dependence of the ^{1}H- and ^{13}C-signal of DSS is similar. However, the pressure dependence of the ^{13}C-signal was previously unknown, but is investigated here.

Table 22 shows the measured ^{13}C chemical shifts of acetamide using the direct referencing approach where the chemical shift of DSS was set to 0 ppm for each pressure and the indirect referencing approach where the ^{1}H resonance frequency of DSS was used together with Ξ = 0.25144953[206] to calculate the ^{13}C chemical shifts.

Table 22: Pressure-dependent experimental ^{13}C chemical shifts of the CH_3 group and the amide group of acetamide using a direct referencing method with δ (^{13}C, DSS) = 0 ppm and an indirect referencing method where the reference was determined by the resonance frequency of ^{1}H DSS and Ξ = 0.25144953[206] following equation (101). The value at 10 bar is used as substitute for ambient conditions due to technical limitations of the pressure pump.

	direct reference				indirect reference			
	CH_3		$C_{carbonyl}$		CH_3		$C_{carbonyl}$	
p/bar	δ / ppm	δ (p)- δ (10 bar) /ppm	δ / ppm	δ (p)- δ (10 bar) /ppm	δ / ppm	δ (p)- δ (10 bar) /ppm	δ / ppm	δ (p)- δ (10 bar) /ppm
10	24.054	0	180.127	0	24.115	0	180.188	0
500	24.048	-0.006	179.999	-0.128	24.264	0.149	180.215	0.027
1000	24.041	-0.013	179.875	-0.252	24.390	0.275	180.239	0.051
1500	24.033	-0.021	179.762	-0.365	24.536	0.421	180.264	0.077
2000	24.026	-0.028	179.654	-0.473	24.659	0.544	180.286	0.098
2500	24.017	-0.037	179.550	-0.576	24.775	0.660	180.308	0.121
3000	24.008	-0.047	179.451	-0.676	24.882	0.767	180.325	0.138

Both referencing methods show a significant difference in the pressure dependence of ^{13}C chemical shifts. Using the direct reference both chemical shifts decrease with pressure but using the indirect reference they both increase. This is due to the fact that DSS shows a significant pressure dependence in the ^{13}C chemical shift but not the ^{1}H chemical shift as previously shown.[7] If the pressure

dependence of the ^{1}H is taken into account as it is done for the indirect referencing method basically no pressure dependence of the standard is included. To further illustrate this point using the indirect referencing method and also being able to measure the ^{13}C DSS signal it is possible to measure the ^{13}C chemical shift of DSS pressure-dependently. Using the reference frequency from (101) the standard definition of the chemical shift from equation (81) can be applied to DSS.

The results are shown in Table 23. The chemical shift was measured with DSS in a sample of benzamide showing a realistic scenario where DSS would be used as an internal reference for another compound.

Table 23: Pressure-dependent ^{13}C chemical shifts of the CH_3 group of DSS using an indirect referencing method where the reference was determined by the resonance frequency of ^{1}H DSS and Ξ = 0.25144953[206] following equation (101). The chemical shift was measured with DSS in a sample of acetamide. The value at 10 bar is used as substitute for ambient conditions due to technical limitations of the pressure pump.

p/bar	δ / ppm	$\delta(p)$- δ (10 bar) /ppm
10	0.061	0
500	0.217	0.156
1000	0.364	0.304
1500	0.503	0.442
2000	0.633	0.572
2500	0.759	0.698
3000	0.875	0.814

Table 23 shows that the ^{13}C chemical shift of DSS shows a large pressure dependence of 0.875 ppm at 3000 bar when referenced to the ^{1}H chemical shift of DSS. The pressure-dependent difference of 0.814 ppm at 3 kbar is also the difference between the direct and indirect reference in Table 22. This shows that the ^{13}C chemical shift of DSS is sensitive to the change in pressure and referencing methods should be consistent among experiment and calculations. Another way of putting this is that the Ξ value that would move the DSS signal back to the reference value at 1 bar is pressure-dependent and defined by

$$\nu_0^{13C,DSS}(p) \overset{!}{=} \Xi(p)\nu_0^{1H,DSS}(p)\,. \tag{102}$$

The pressure-dependent Ξ values that fulfill this equation are given in Table 24.

Table 24: Pressure-dependent Ξ values such that the ^{13}C chemical shift of the CH_3 group of DSS remains at ca. 0.06 ppm with respect to the resonance frequency of ^{1}H DSS and a Ξ = 0.25144953[206] at ambient conditions. The chemical shift was measured with DSS in a sample of acetamide. Values are rounded to the same number of digits as reported in literature. The pressure-dependent internal standard of ^{1}H DSS was used. The value at 10 bar is used as substitute for ambient conditions due to technical limitations of the pressure pump.

p/bar	Ξ^{13C}
10	0.25144953
500	0.25144957
1000	0.25144960
1500	0.25144964
2000	0.25144967
2500	0.25144970
3000	0.25144973

Table 24 shows that the Ξ-values only need very small adjustments to remove the pressure dependence of the ^{13}C chemical shift from DSS. With theses values the inconsistency between referencing methods can also be restored, whenever it is not possible to measure the reference frequency for ^{13}C DSS.

4.3.1 Summary and outlook

Experimental studies showed that pressure-dependent chemical shifts measurements are extremely reliable even though the pressure-dependence of chemical shifts might be small. Analysis of multiple measurements of the same compound showed that intrinsic errors even for different samples and among different spectrometers are less than 0.03 ppm for ^{1}H and the error of pressure-dependent chemical shift change is even less. With this result in mind, NMR measurements are a valuable benchmark for theoretical calculations and will be used in this way within this thesis.

A different picture emerges for ^{13}C for which a significant uncertainty emerges from the choice of reference. The ^{13}C signal of DSS used as a reference at ambient conditions turned out to be the most pressure-sensitive. Whenever DSS is still used a reference special considerations regarding the referencing method are needed, otherwise large inconstancies can occur. Using internal DSS as a direct reference for ^{13}C and using the pressure-dependent ^{1}H signal together with a pressure-independent Ξ-value here yields a difference of ca. 0.8 ppm. This inverts trends for the pressure-dependent chemical shift change for all measured compounds. However, new pressure-dependent Ξ-values to remove this reference artifact have also been proposed and are recommended for all future indirect references of pressure-dependent ^{13}C-NMR measurements.

From a benchmark point of view the direct referencing most closely resembles the theoretical approach, which is why within this thesis pressure-dependent chemical shifts from calculations will

use a direct reference for calculated shielding constants and experimental benchmark data will use the same method based on equation (81). Previous pressure-dependent ^{13}C chemical shift calculations[7] which used inconsistent referencing method for measured and calculated chemical shifts relied, most likely, on significant error cancellation between calculations and measurements to still match trends. A deeper analysis of pressure-dependent NMR calculations will follow in the next chapter.

For referencing ^{13}C chemical shifts at ambient conditions both referencing methods agree up to a difference of only 0.06 ppm which justifies both referencing methods and is the reason why both of them were established.

4.4 Pressure-dependent NMR calculations

The goal of this chapter is to extend on previous, both theoretical and experimental, findings and combine highly accurate NMR predictions with thermodynamics. This is done by first analyzing previously developed methods for NMR calculations, namely FFMD-based methods using no solvation (VD) and the "ref, QM/MM" solvation method from chapter 4.2 and single point calculations using both PCM optimized geometries and newly developed EC-RISM geometries. The analysis is performed at ambient conditions and for the pressure-dependent chemical shift changes at 3 kbar. As a benchmark dataset a set of conformationally inflexible amides is used to probe the applicability of these methods without needing to account for conformational or tautomeric changes. As chapter 4.1.2 showed that structural changes need to be included for the amide proton of *cis*-NMA a closer look at this is taken for more nuclei. The expectation based on this is that EC-RISM geometries are an improvement over PCM geometries. In comparison to the "$ref_{QM/MM}$" solvation method it is expected based on Table 11 and Table 13 that the prediction performance is similar for the single-point method, at least using PCM geometries. With the experimental investigation of the ^{13}C reference from chapter 4.3 this will be further investigated for theoretical methods. The initial approach however is to be methodologically consistent and directly reference both experimental and theoretical ^{13}C chemical shifts.

In the second part of this chapter the learnings from the initial pressure-dependent benchmark calculations are used to answer the question how pressure sensitive the protein backbone is. This is done for the two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe as model systems for larger peptides and proteins for which experimental data was measured by M. Beck-Erlach. Here direct energetic calculations by T. Pongratz [224] are supplemented by an NMR fit that uses both experimental and calculated data generated as part of this thesis. It is applied to NMR calculations for the two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe which were previously also performed by T. Pongratz.[224] In the end a thermodynamic fit, co-developed with T. Pongratz and S. Kast, uses the corrected single point NMR calculations on the amide protons of Ac-Ala-NHMe and Ac-Gly-NHMe to fit the conformational equilibrium to match experimental chemical shifts. This shows an application which combines both spectroscopy and thermodynamics but requires highly accurate chemical shifts, for which methods are developed and analyzed here.

4.4.1 Pressure-dependent benchmark calculations on an amide subset

Now taking all developed methods and insights from previous chapters into account, an amide subset was used as benchmark set for pressure-dependent geometry optimizations and to compare single point vs. ensemble-based NMR calculations. The molecules of this subset are shown in Figure 24.

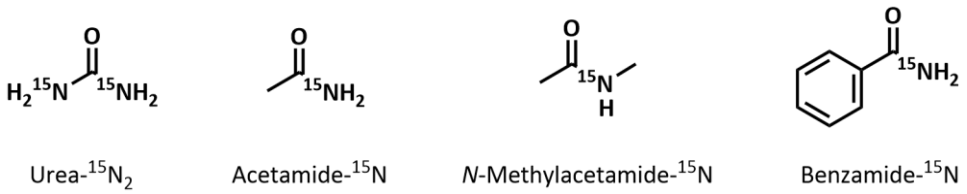

Figure 24: Molecules of the amide subset. For all molecules except urea the amide protons show distinguishable NMR signals for the *cis*- and *trans*-conformation. In total 13 distinguishable protons, 11 distinguishable carbons and 4 distinguishable nitrogens are present.

Here ambient condition calculations are compared in Figure 25, **A**, **C**, **E** for ^{1}H, ^{13}C and ^{15}N respectively. Second the pressure-dependent chemical shift change is shown in Figure 25, **B**, **D**, **F**. Here the difference in chemical shifts at 3 kbar as the highest experimentally available pressure against ambient conditions are compared. The raw data for these is shown in Table 25. Error metrices are compared in Table 26 for ambient conditions and the pressure-dependent changes.

From this overview over individual datapoints and overall metrics the key takeaways are the following.

For ambient conditions EC-RISM geometries overall show the best results with an RMSE of 0.76 ppm for ^{1}H, 2.37 ppm for ^{13}C and 5.02 ppm for ^{15}N. In general, single point calculations consistently outperform the ensemble-based FFMD methods, especially for ^{13}C and ^{15}N with EC-RISM geometries showing significant improvements compared to PCM-optimized geometries throughout all calculations at ambient conditions. The higher level of theory with EC-RISM compared to PCM yields more accurate structures, which was previously unknown due to the new development of EC-RISM geometry optimizations.

Looking into this in more detail, the full solvation method consistently yields the best amide chemical shift. This adds to the previous finding of chapter 4.2 where the the “ref,QM/MM” solvation method was not performed for all model systems, but similar methods like “ECR,full” (Table 16) were comparable to single point calculations (Table 13). The “$FFMD_{ref,QM/MM}$” method suffers from worse predictions of the aromatic protons and overall large errors for almost all ^{13}C and ^{15}N chemical shifts. Here single point calculations yielded close agreement with experiments in line with the findings from Table 14 were the nuclei of NMA were also well predicted by the “opt,ECR,0” method which is the same as the here called “PCM” method.

Overall, at ambient conditions the use of EC-RISM geometries is recommended. The “$FFMD_{ref,QM/MM}$” method is only recommended for the amide protons. A further investigation needs to be made into the specific parmeters used especially for carbonyl carbon, aromatic protons and, as shown in Table 27, ammonia as the ^{15}N reference.

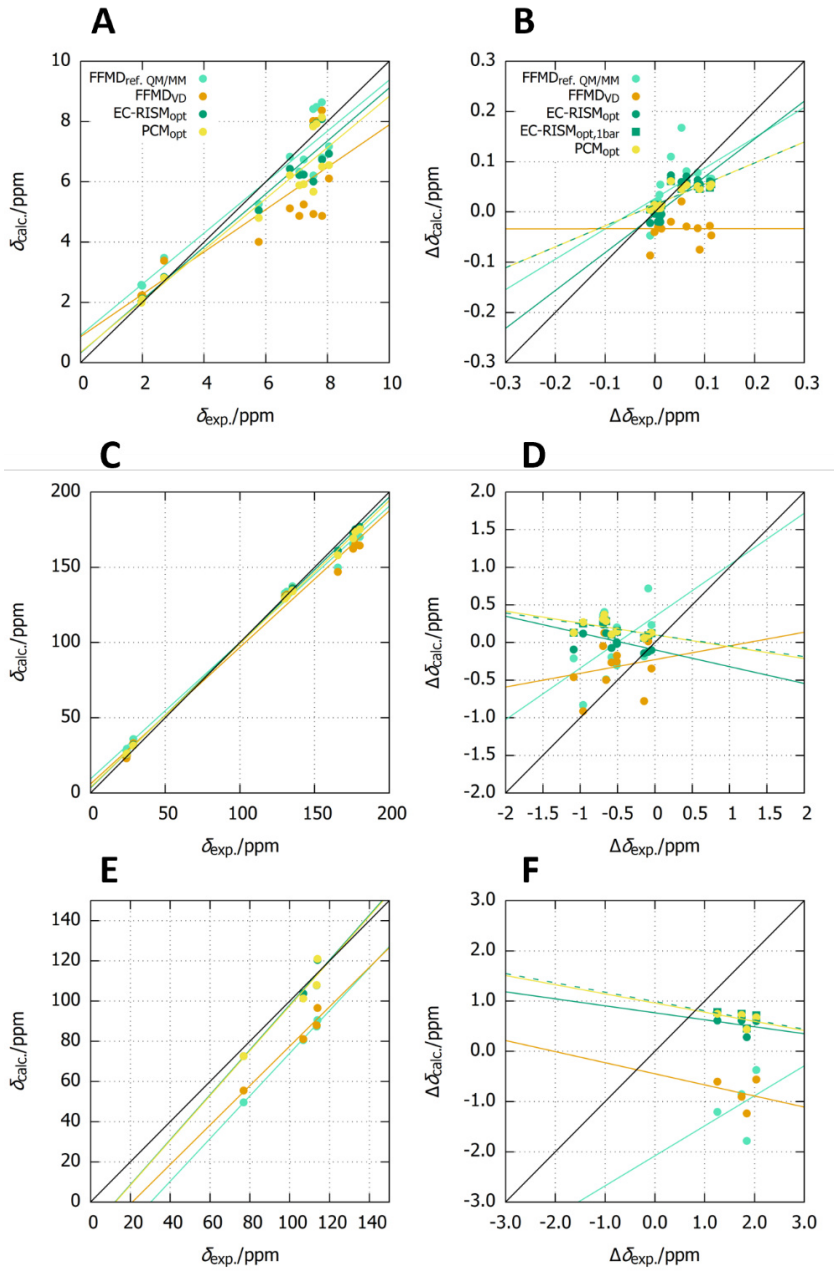

Figure 25: Calculated vs. experimental chemical shifts at ambient conditions for all ^{1}H (**A**), ^{13}C (**C**) and ^{15}N (**E**) atoms of the amide subset. Corresponding chemical shift differences from 3 kbar to ambient conditions in **B** (^{1}H), **D** (^{13}C) and **F** (^{15}N). Calculation methods include PCM- and EC-RISM-optimized single structure calculations and FFMD based ensemble calculations. For the ensembles calculations 400 snapshots were taken equidistantly and chemical shifts were calculated including no solvation (VD) or the "ref QM/MM" solvation method. For EC-RISM-optimized structures pressure-dependent geometry optimizations are distinguished from pressure independent geometry optimizations (opt,1bar) shown with squares and dashed lines. For geometry optimizations the B3LYP/6-311+G** level of theory was employed. For all NMR calculations the OLYP/6-311+G** level of theory was used. ^{1}H and ^{13}C chemical shifts are directly referenced against the methyl group of DSS. Calculated ^{15}N chemical shifts are referenced against aqueous NH_3 with a chemical shift of -19.4 ppm against the experimental reference liquid NH_3.[266] Ideal predictions would follow the black diagonal. Raw data of this figure is shown in Table 25. Raw data for the calculations (structures with corresponding shielding constants) is given in the electronic supplementary information, see chapter 6.1.

Table 25: Raw data for the calculated chemical shifts with respect to experimental data and pressure-dependent chemical shift differences with respect to experimental data for the 13 distinguishable protons, 11 distinguishable carbons and 4 distinguishable nitrogens of the amide subset. Experimental chemical shifts are referenced directly to DSS at each pressure for ^{1}H and ^{13}C. Experimental ^{15}N are referenced against ^{1}H DSS at each pressure with Ξ = 0.25144953.[206] Chemical shifts and chemical shift differences are depicted in Figure 25. The best calculation for each nucleus is highlighted in bold. Raw data for all experimental chemical shift is shown in the appendix (Table 48 to Table 52). Raw data for the calculations (structures with corresponding shielding constants) and experimental spectra are given in the electronic supplementary information, see chapter 6.1

	1 bar					3 kbar – 1 bar					
nucleus	δ_{exp}	$\Delta\delta_{VD}$	$\Delta\delta_{Full}$	$\Delta\delta_{EC\text{-}RISM}$	$\Delta\delta_{PCM}$	$\Delta\delta_{exp}$	$\Delta\Delta\delta_{VD}$	$\Delta\Delta\delta_{Full}$	$\Delta\Delta\delta_{EC\text{-}RISM}$	$\Delta\Delta\delta_{EC\text{-}RISM\ (1bar)}$	$\Delta\Delta\delta_{PCM}$
urea											
N**H**	5.770	-1.763	**-0.506**	-0.721	-0.965	0.089	**0.000**	-0.165	-0.034	-0.044	-0.044
CO	165.50	-18.50	-15.65	**-4.45**	-7.51	-0.66	**0.16**	0.17	0.78	0.94	0.94
NH	76.84	-21.39	-27.25	**-4.25**	-4.29	1.85	-3.09	-3.63	-1.57	**-1.39**	-1.41
acetamide											
H-me.	1.993	0.247	0.556	0.132	**0.108**	0.006	-0.037	0.016	-0.026	**0.001**	**0.001**
N**H**-trans	7.543	-2.426	**-0.712**	-1.114	-1.327	0.054	-0.100	0.013	0.006	0.002	**0.001**
N**H**-cis	6.782	-1.848	**-0.571**	-0.779	-1.112	0.114	-0.093	0.054	**-0.054**	-0.068	-0.068
C-me.	24.05	-0.94	3.90	**1.63**	1.72	-0.05	-0.30	0.28	**-0.05**	0.18	0.18
C-carb.	180.13	-15.66	-9.97	**-3.26**	-4.87	-0.68	**0.81**	1.08	1.00	1.03	1.05
NH	112.15	-24.24	-24.88	**-4.30**	-4.24	2.07	-2.63	-2.44	-1.47	**-1.35**	-1.39
NMA											
H-me.-1	1.967	0.257	0.613	0.050	**0.019**	0.010	-0.015	0.044	-0.030	**-0.003**	**-0.003**
N**H**-trans	7.817	-2.951	**-1.012**	-1.078	-1.313	0.029	-0.049	0.081	0.044	**0.032**	**0.032**
H-me.-2	2.702	0.678	0.776	0.137	**0.111**	-0.011	-0.076	-0.036	**-0.011**	0.016	0.016
N**H**-cis	7.086	-2.220	**-0.748**	-0.867	-1.200	0.084	-0.116	**-0.006**	-0.020	-0.033	-0.032
C-me.-1	24.39	0.69	4.96	**1.85**	1.92	-0.09	0.11	0.81	**-0.04**	0.19	0.19
C-carb.	177.35	-11.87	-6.47	**-2.19**	-3.67	-0.69	**0.64**	1.06	0.93	0.98	1.00
C-me.-2	28.70	4.41	7.15	**3.31**	2.82	-0.14	-0.64	-0.05	**0.00**	0.20	0.20
NH	112.63	-16.04	-22.11	**7.81**	8.35	1.20	-1.80	-2.41	-0.59	**-0.42**	-0.45
benzamide											
H-para	7.639	0.372	0.845	0.306	**0.282**	0.013	-0.046	-0.018	-0.018	**0.000**	0.001
H-meta	7.534	0.483	0.883	0.322	**0.301**	0.009	-0.039	0.025	-0.020	**0.003**	**0.003**
H-ortho	7.813	0.553	0.823	**0.267**	0.326	-0.001	-0.039	0.019	**-0.003**	0.015	0.015
N**H**-trans	8.050	-1.943	**-0.875**	-1.118	-1.501	0.063	-0.092	0.018	**0.007**	**-0.007**	-0.010
N**H**-cis	7.227	-1.979	**-0.485**	-0.993	-1.304	0.110	-0.138	**-0.045**	-0.049	-0.062	-0.060
C-para	135.24	-1.14	2.09	**-0.49**	-0.61	-0.52	0.27	**0.22**	0.52	0.66	0.67
C-meta	131.45	0.51	2.34	**-0.41**	-0.77	-0.58	0.32	**0.39**	0.51	0.68	0.69
C-ortho	130.09	0.87	2.72	**-0.43**	-0.96	-0.50	**0.33**	0.70	0.48	0.65	0.65
C-ipso	135.34	0.75	**-0.50**	0.55	-1.31	-1.09	**0.63**	0.88	1.00	1.22	1.22
C-carb.	175.90	-13.64	-10.49	**-2.94**	-6.72	-0.96	0.05	**0.13**	1.08	1.21	1.23
NH2	106.77	-25.65	-26.23	**-3.02**	-5.47	1.74	-2.65	-2.59	-1.12	**-0.99**	-1.01

Table 26: Root mean squared error (RMSE) mean absolute error (MAE) and mean signed error (MSE) of the ^{1}H, ^{13}C and ^{15}N for the chemical shifts at 1 bar and the chemical shift differences at 3 kbar compared to experiments for the 13 distinguishable protons, 11 distinguishable carbons and 4 distinguishable nitrogens of the amide subset. Regression parameters m and b are given for the linear regression together with the R^2 coefficient. Calculation methods include PCM- and EC-RISM-optimized single structure calculations and FFMD based ensemble calculations. For the ensembles calculations 400 snapshots were taken equidistantly and chemical shifts were calculated including no solvation (VD) or "ref,QM/MM" solvation. For geometry optimizations the B3LYP/6-311+G** level of theory was employed. For all NMR calculations the OLYP/6-311+G** level of theory was used. ^{1}H and ^{13}C chemical shifts are directly referenced against the methyl group of DSS. 15N chemical shifts are referenced against aqueous NH_3 with a chemical shift of -19.4 ppm against the experimental reference liquid NH_3.[266] The best performing model for each metric is highlighted in bold.

Nucleus	Method	RMSE	MAE	MSE	m	b	R^2
			$\Delta\delta_{1bar}$				
^{1}H	$FFMD_{VD}$	1.648	1.363	-0.965	0.703	0.864	0.645
	$FFMD_{ref,QM/MM}$	0.788	0.731	-0.032	0.846	0.914	0.876
	$EC\text{-}RISM_{opt}$	**0.758**	**0.607**	**-0.420**	**0.879**	**0.324**	**0.922**
	PCM_{opt}	0.968	0.760	-0.583	0.849	0.345	0.881
^{13}C	$FFMD_{VD}$	9.23	6.27	-4.95	0.907	6.058	0.990
	$FFMD_{ref,QM/MM}$	7.44	6.02	-1.81	0.906	9.376	0.993
	$EC\text{-}RISM_{opt}$	**2.37**	**1.96**	**-0.62**	**0.966**	**3.414**	**0.999**
	PCM_{opt}	3.78	2.99	-1.81	0.952	3.895	0.999
^{15}N	$FFMD_{VD}$	22.78	22.52	-22.52	**0.981**	-20.567	0.951
	$FFMD_{ref,QM/MM}$	25.85	25.81	-25.81	1.061	-32.105	**0.996**
	$EC\text{-}RISM_{opt}$	**5.02**	**4.84**	**-1.63**	1.118	**-13.732**	0.938
	PCM_{opt}	5.66	5.58	-2.10	1.115	-13.960	0.922
			$\Delta\Delta\delta_{3\ kbar-\ 1\ bar}$				
^{1}H	$FFMD_{VD}$	0.092	0.078	-0.078	0.001	-0.033	0.000
	$FFMD_{ref,QM/MM}$	0.048	0.039	**0.009**	0.604	0.027	0.250
	$EC\text{-}RISM_{opt}$	**0.030**	0.025	-0.017	**0.754**	**-0.006**	0.677
	$EC\text{-}RISM_{opt,1\ bar}$	0.031	0.023	-0.014	0.399	0.013	0.679
	PCM_{opt}	**0.030**	**0.022**	-0.012	0.417	0.014	**0.681**
^{13}C	$FFMD_{VD}$	**0.45**	**0.38**	**0.21**	0.182	-0.227	0.038
	$FFMD_{ref,QM/MM}$	0.64	0.52	0.52	**0.686**	0.346	**0.263**
	$EC\text{-}RISM_{opt}$	0.70	0.58	0.57	-0.223	**-0.098**	0.230
	$EC\text{-}RISM_{opt,1\ bar}$	0.81	0.72	0.72	-0.145	0.102	0.262
	PCM_{opt}	0.82	0.73	0.73	-0.157	0.102	0.262
^{15}N	$FFMD_{VD}$	2.58	2.55	-2.55	-0.221	**-0.448**	0.055
	$FFMD_{ref,QM/MM}$	2.82	2.77	-2.77	**0.597**	-2.080	0.113
	$EC\text{-}RISM_{opt}$	1.24	1.19	-1.19	-0.139	0.767	0.079
	$EC\text{-}RISM_{opt,1\ bar}$	**1.11**	**1.04**	**-1.04**	-0.184	0.992	0.172
	PCM_{opt}	1.13	1.07	-1.07	-0.182	0.961	**0.182**

Table 27: Reference shielding constants / ppm for the calculation of the ^{1}H, ^{13}C and ^{15}N chemical shifts at 1 bar and difference at 3 kbar. Calculation methods include PCM- and EC-RISM-optimized single structure calculations and FFMD based ensemble calculations. For the ensembles calculations 400 snapshots were taken equidistantly and chemical shifts were calculated including no solvation (VD) or the "ref,QM/MM" solvation method. For EC-RISM-optimized structures pressure-dependent geometry optimizations are distinguished from pressure independent geometry optimizations using the same structure optimized at 1 bar. For all geometry optimizations the B3LYP/6-311+G** level of theory was employed. For all NMR calculations the OLYP/6-311+G** level of theory was used. As a reference $AIMD_{ref,QM/MM}$ from Table 10 are given too for 1 bar.

Nucleus	Method	σ_{1bar}	$\Delta\sigma_{3kbar-1bar}$
^{1}H	$FFMD_{VD}$	32.1211	-0.0326
	$FFMD_{ref,QM/MM}$	31.9512	-0.0218
	$EC\text{-}RISM_{opt}$	31.9284	0.0022
	$EC\text{-}RISM_{opt,1\ bar}$	31.9284	0.0022
	PCM_{opt}	31.8398	0.0021
	$AIMD_{ref,QM/MM}$	31.18	-
^{13}C	$FFMD_{VD}$	187.996	-0.280
	$FFMD_{ref,QM/MM}$	187.412	-0.245
	$EC\text{-}RISM_{opt}$	189.182	0.102
	$EC\text{-}RISM_{opt,1\ bar}$	189.182	0.102
	PCM_{opt}	188.292	0.101
	$AIMD_{ref,QM/MM}$	182.33	-
^{15}N	$FFMD_{VD}$	238.811	-0.703
	$FFMD_{ref,QM/MM}$	213.938	-2.152
	$EC\text{-}RISM_{opt}$	244.608	0.295
	$EC\text{-}RISM_{opt,1\ bar}$	244.608	0.345
	PCM_{opt}	245.149	0.339

Moving over to the chemical shift differences at 3 kbar a totally different picture emerges. Overall, only the ^{1}H chemical shift differences are predicted reliably with an R^2 of more than 0.6 for the single point calculations (opt). Here the uncertainty of the ensemble-based methods play a large role, as the uncertainty of a chemical shift is on the order of 0.1 ppm for ^{1}H and 0.2 ppm for ^{13}C based on the findings of chapter 4.2 which is especially significant compared to the experimental range of ^{1}H. Accurate ^{1}H chemical shift differences would likely require a significantly larger sample size than 400 snapshots.

However, the "$FFMD_{ref,QM/MM}$" method is still noteworthy as it is the only one yielding a positive slope for ^{13}C and ^{15}N making this the only reliable and therefore method for accurate relative chemical shift differences of ^{13}C and ^{15}N. In terms of absolute values the "$FFMD_{ref,QM/MM}$" method suffers from a large change of the ^{15}N reference shieldings (2.15 ppm, see Table 27). The experimental data was measured using an indirect reference which basically includes no pressure dependence. Removing the pressure dependence of the reference would give the "$FFMD_{ref,QM/MM}$" method also the best predictions (MAE

ca. 0.6 ppm). However it is not clear if this is an issue of the experimental data, as found for ^{13}C in chapter 4.4 or a false description of ammonia using the "$FFMD_{ref,QM/MM}$" method. This would require a study of ammonia as internal standard similar to the study of DSS in chapter 4.4. The fact that a consistent referencing method is necessary can also be seen for ^{13}C. Here the experimental chemical shift increase of DSS, shown in Table 23, is only shown by the FFMD-based methods, even though to a smaller extent, (ca. 0.3 ppm compared to ca. 0.8 ppm) showing that all methods are not sufficiently accurate in the description of DSS. The fact that the FFMD-based methods at least yield the right trend compared to experiments plays a large part as to why these methods also show the smallest RMSE for the ^{13}C chemical shift differences. All methods expect "$FFMD_{VD}$" overestimate the ^{13}C chemical shift differences for all nuclei also due to the overestimation of the reference shieldings at 3 kbar. Using the experimental insights from chapter 4.3 this error source however can be clearly quantified and further improvements in calculations based on reliable experiment reference data are possible. An interesting approach would be to use AIMD simulations of DSS at high pressure and study the intra- and intermolecular structure similar to chapter 4.2.

To answer another research question of this study, namely disentangling the structural and solvation part of the chemical shift differences under high pressure, the solvation part clearly plays the dominant role in the accurate prediction of chemical shift differences. This can be seen from a couple of observations. The "$FFMD_{VD}$" method, only showing the structural part, shows almost no correlation for the chemical shift differences (R^2 values of 0, 0.038 and 0.055). Only after inclusion of the solvent in the "$FFMD_{ref,QM/MM}$" method, better correlation occurs. Secondly the difference between "$EC\text{-}RISM_{opt}$" and "$EC\text{-}RISM_{opt,1bar}$" is small throughout all metrics. While the addition of the pressure-dependent structural change with "$EC\text{-}RISM_{opt}$" especially improves the *cis*-protons as shown before for NMA the overall improvement is small with the ^{15}N predictions even being worse.

Based on this first benchmark study and earlier finding a few recommendations for further studies to improve the prediction quality can be made.

An MLMD simulation or even just the MLMD refinement of FFMD ensembles could improve predictions using the "ref,QM/MM" solvation method as seen in chapter 4.2.4 and 4.2.5, however it is unclear how accurately these models perform under high-pressure conditions. Especially for single point calculations varying the level of theory for both the geometry optimization as well as the NMR calculation could further improve the predictions. This was untouched during this benchmark study but could have significant impacts. One goal of this would be to improve the pressure dependence of ^{13}C-DSS which showed the wrong trend compared to experimental measurements.

4.4.2 Calculation of conformational equilibria under high pressure from chemical shifts

Based on the findings of the previous chapter on how to calculate accurate pressure-dependent chemical shifts, it is now possible to apply this to a completely different question, the conformational equilibria under high pressure. This study focuses on the two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe as model systems for larger peptides and proteins. For these pressure-dependent NMR measurements can not distinguish individual conformations, as the conformational change is faster than the NMR measurement. This is in contrast to the model system of NMA where *cis*- and *trans*-conformations can be distinguished as was done in the creation of the amide subset within this thesis.

Previous work by T. Pongratz[224] showed that the conformational equilibria of NMA could be well reproduced using high-level QM calculations of the free energy in solution. With the same methods the conformational equilibrium of Ac-Ala-NHMe and Ac-Gly-NHMe is predicted to be pressure insensitive with a maximum change of less than 0.3 kcal mol^{-1} at 5 kbar in line with results for NMA which closely matched experimental observations.[224]

The goal is to now use an additional, independent method to further validate the findings from the energy calculations, namely fitting of calculated chemical shifts for each conformer to experimental chemical shifts which results as an average over conformers.

To outline the theoretical framework (developed with T. Pongratz and S. Kast) taking the two most common conformers from energy calculations (P_{II} and α) the chemical experimental chemical shift can be fitted using:[224]

$$\delta_{\mathrm{corr}}^{\mathrm{mean}}(p)=\frac{1}{1+\exp\left(-\beta\Delta G_0(p_0)+s(p-p_0)+t(p-p_0)^2\right)}\delta_{\mathrm{corr}}^{\mathrm{i}} + \left(1-\frac{1}{1+\exp\left(-\beta\Delta G_0(p_0)+s(p-p_0)+t(p-p_0)^2\right)}\right)\delta_{\mathrm{corr}}^{\mathrm{j}} \tag{103}$$

where ΔG_0, s and t are fitted to reproduce the pressure-dependent change in population in terms of the free energy difference.

This fitting however requires accurate chemical shift predicitions which, as shown before, can be difficult to achieve. To overcome this problem a fit of chemical shift is performed for the amide subset, treated as an independent training dataset, to then apply the fitting parameters to the amide protons of Ac-Ala-NHMe and Ac-Gly-NHMe.

To match the level of theory for the initial NMR calculations of Ac-Ala-NHMe and Ac-Gly-NHMe MP2/6-311+G(d,p)/EC-RISM/PSE-3 was used throughout this fitting. In addition, both pressure-independent

PCM geometries and pressure-dependent EC-RISM geometries. Both methods showed among the best performances for the amide protons (see Table 25) and do not require the huge computational effort for the ensemble-based calculations that are unfeasible at the MP2-level.

The fit for the amide protons (co-developed with T. Pongratz and S. Kast) then has the form:

$$\delta_{corr}(p) = \delta_{orig}(p) + a(p - p_0)^2 + b(p - p_0) + \delta_0 \tag{104}$$

Here, δ_{corr} is the corrected chemical shift which is corrected by the scaling parameter for a quadratic term a, and a linear parameter b. δ_0 is an additive offset for the chemical shift.

As both PCM and EC-RISM geometries were used they were investigated and fitted separately. Resulting uncorrected and corrected chemical shifts are shown in Figure 26 with the fit parameters listed in Table 28.

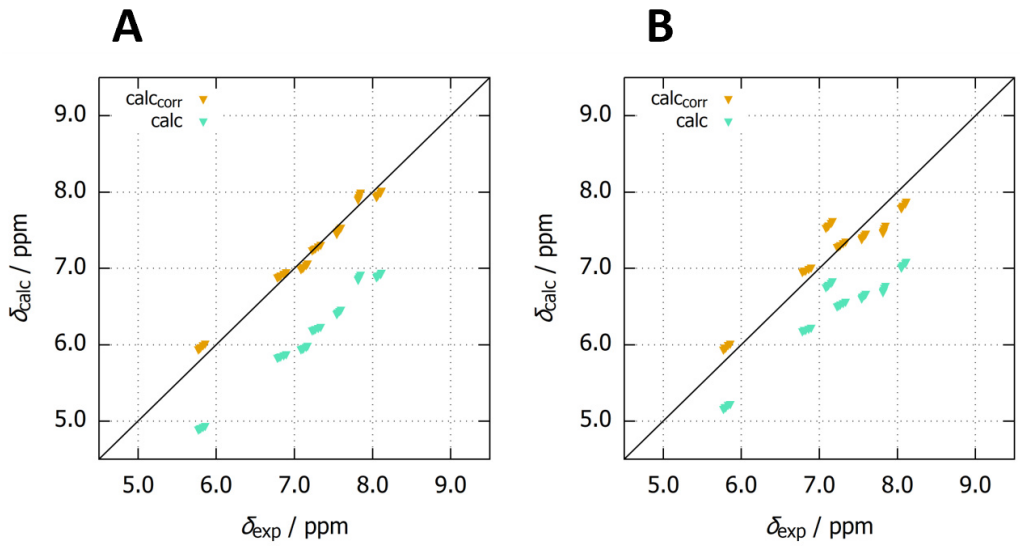

Figure 26: Uncorrected (blue triangles) and corrected (yellow triangles) pressure-dependent chemical shifts of the seven distinguishable amide protons of the amide subset. Panel **A** shows the results for the B3LYP/6-311+G(d,p)/PCM optimized structures, whereas in **B** the corresponding results for the B3LYP/6-311+G(d,p)/EC-RISM results are shown. Chemical shifts were calculated on the MP2/6-311+G(d,p)/EC-RISM/PSE-3 level of theory and are directly referenced against DSS calculated on the same level of theory but using only PCM geometries.[8] Calculations were performed at 1, 500, 1000, 2000 and 3000 bar. Raw data is published in the electronic supplementary material, see chapter 6.1.

Table 28: Fit parameters for the amide subset following equation (104) for PCM- and pressure-dependent EC-RISM-geometries. NMR calculations were performed on the MP2/6-311+G(d,p)/EC-RISM/PSE-3 level of theory.

Parameter	PCM geometries	EC-RISM geometries
δ_0 / ppm	1.0516	0.7766
b / ppm·kbar^{-1}	0.0089	0.0024
a / ppm·kbar^{-2}	-0.0002	0.0006

Figure 26 shows that using the fit from equation (104) with the parameters from Table 28 chemical shifts can be matched very closely to experimental chemical shifts with mean absolute errors of 0.09 ppm and 0.21 ppm for PCM and EC-RISM optimized geometries respectively averaged over all nuclei and pressures. It is noteworthy that the pressure-dependent parameters a and b are very small compared to the intrinsic error δ_0. This is in line with the pressure dependence of the chemical shift itself which was found to be less than 0.1 ppm·kbar^{-1}. The linear parameter b has a much bigger impact compared to the quadratic parameter a and it is much smaller for EC-RISM geometries which better reproduce both the intrinsic chemical shifts and the pressure-dependent trend.

To illustrate this for one example, the chemical shift difference of *cis*-NMA is shown with and without correction for both sets of geometries in Figure 33 in comparison to experiments.

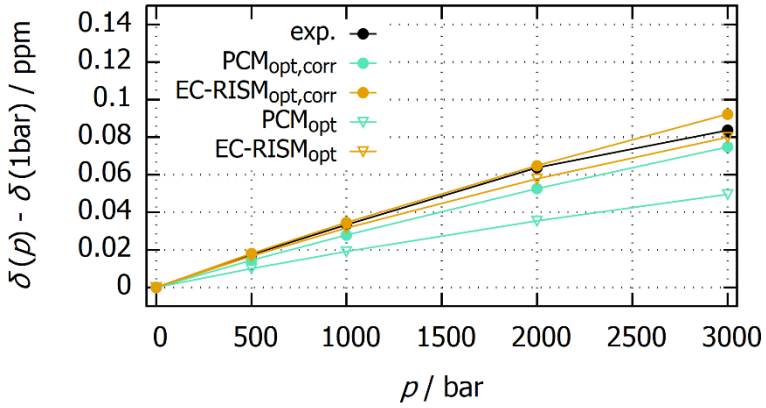

Figure 27: Calculated and experimental pressure-dependent chemical shift differences with respect to 1 bar for the amide proton of *cis*- NMA. Calculations include results for the B3LYP/6-311+G(d,p)/PCM (PCM_{opt}) and B3LYP/6-311+G(d,p)/EC-RISM ($EC\text{-}RISM_{opt}$) structures of theory and are directly referenced against DSS calculated on the same level of theory but using only PCM geometries.[8] Both methods are also shown with the fit parameters from Table 28 applied ("corr").

Figure 28 shows the impact of the fit parameters *a* and *b* from Table 28. For PCM the relatively large *b* yields a significant, but mostly linear, increase in the pressure dependence. For EC-RISM this is not as significant, due to the smaller linear correction. The overall shape following the quadratic equation is already given for the raw data, and does not require a large change to match the experimental curve. This explains the small *a* values for both methods. In the case of *cis*-NMA the correction yields a very close match to experiments for the chemical shift increase for both methods.

As both the PCM-geometries and EC-RISM-geometries now have the capability to accurately reproduce experimental amide chemical shifts with fitting, both were investigated on their accuracy for predicting dipeptide chemical shifts. Here initial calculations performed by T. Pongratz[224] were corrected using the parameters from Table 28. The resulting chemical shifts are shown in Figure 33 for PCM optimized structures and in Figure 29 for EC-RISM optimized structures. Here both the individual chemical shifts are given as well as the population-weighted average “mean” with the directly calculated (“calc”) and the fitted populations (“fit”).

Figure 33 and Figure 29 show a clear improvement of especially the absolute chemical shifts due to the fitting. The absolute chemical shift is still mostly underestimated, however the error gets significantly smaller. Comparing both PCM and EC-RISM geometries shows that the error after applying the correction is similar for both methods. This means that the correction is transferable to other amides as done in this case. It can be seen that for EC-RISM structures (Figure 29) the uncorrected chemical shifts show a smaller error compared to PCM structures (Figure 33). The larger correction from the PCM optimized structures (Table 28) then corrects for this and corrected chemical shifts are close for both methods.

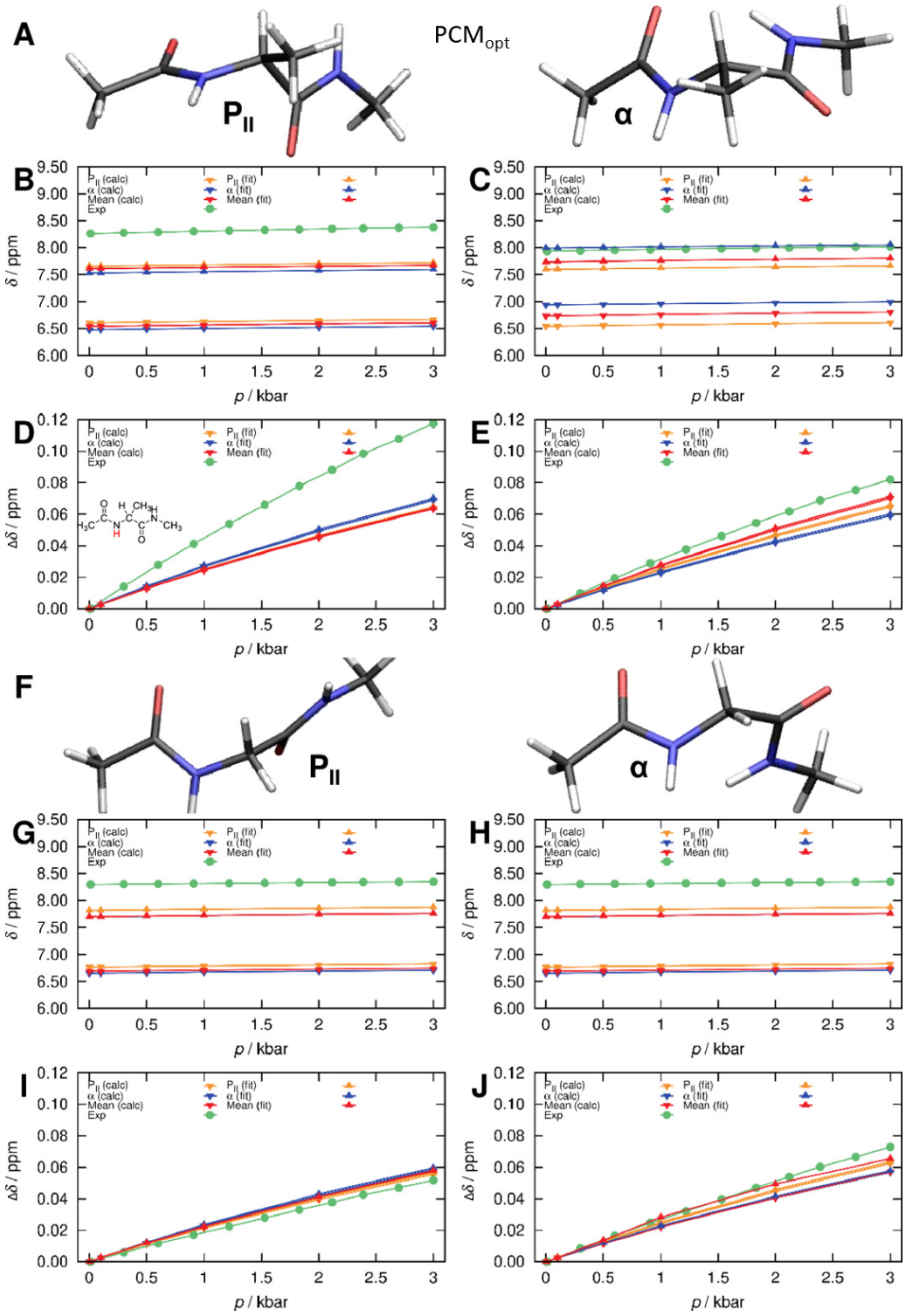

Figure 28: Pressure-dependent chemical shifts of Ac-Ala-NHMe and Ac-Gly-NHMe for the two main conformers P_{II} and α. Panels A (Ac-Ala-NHMe) and F (Ac-Gly-NHMe) show the three-dimensional structures. In panels B and C the chemical shifts of the two amide protons of Ac-Ala-NHMe are shown and the corresponding pressure-dependent changes are depicted in panels D and E. The absolute chemical shifts of Ac-Gly-NHMe are depicted in panels G and H, and the pressure-changes in I and J. The downward triangles show the pure MP2/6-311+G(d,p)/EC-RISM//B3LYP/6-311+G(d,p)/PCM results, whereas for the upward triangles the chemical shift correction scheme from Table 28 is applied. Figure created and calculations performed by T.Pongratz.[224]

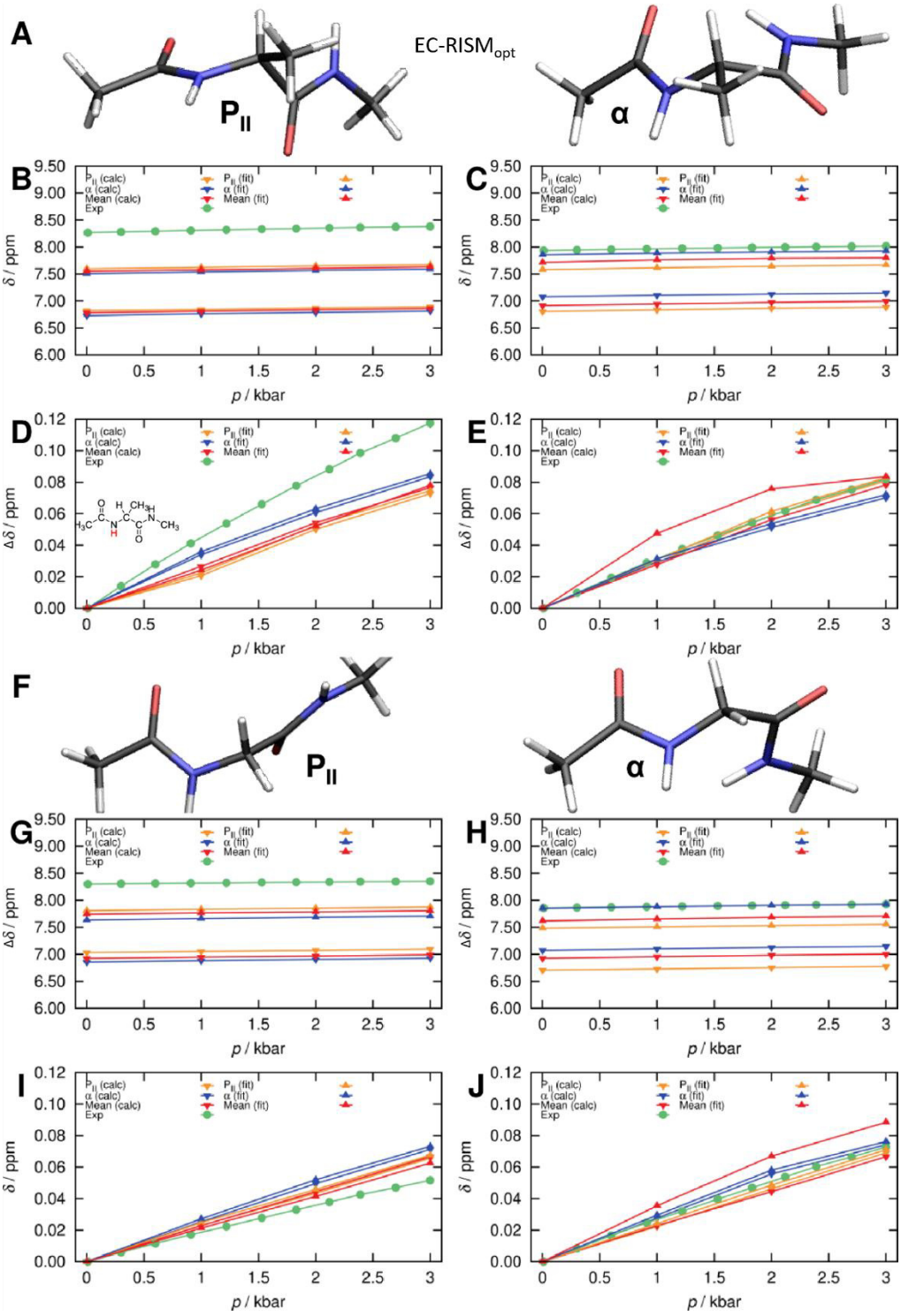

Figure 29: Pressure-dependent chemical shifts of Ac-Ala-NHMe and Ac-Gly-NHMe for the two main conformers P_{II} and α. Panels A (Ac-Ala-NHMe) and F (Ac-Gly-NHMe) show the three-dimensional structures. In panels B and C the chemical shifts of the two amide protons of Ac-Ala-NHMe are shown and the corresponding pressure-dependent changes are depicted in panels D and E. The absolute chemical shifts of Ac-Gly-NHMe are depicted in panels G and H, and the pressure-changes in I and J. The downward triangles show the pure MP2/6-311+G(d,p)/EC-RISM//B3LYP/6-311+G(d,p)/EC-RISM results, whereas for the upward triangles the chemical shift correction scheme from Table 28 is applied. Figure created and calculations performed by T.Pongratz.[224]

With the corrected chemical shift, it is now possible to calculate the pressure dependent populations according to equation (103). The resulting parameters are shown in Table 29, the populations in comparison to direct calculations is given in Table 30. For the comparison two different methods for the pressure-dependent free energy calculation are given with the first one taking the pressure-dependent free energy from an EC-RISM calculation and the second one called the PMV-integration according to: [224]

$$G(p) = G^0 + \int_{p^0}^{p} V_m(p)\, dp' \tag{105}$$

where $G(p)$ is the then molar free energy resulting from partial molar volume.

Table 29: Fit parameters ΔG_0, *s* and *t* for Ac-Ala-NHMe and Ac-Gly-NHMe following equation (103) for PCM- and pressure-dependent EC-RISM-geometries. NMR calculations were performed on the MP2/6-311+G(d,p)/EC-RISM/PSE-3 level of theory with geometries optimized using B3LYP/6-311+G(d,p). ΔG_0 is in kcal mol^{-1}, s is in kcal mol^{-1} $kbar^{-1}$ and t given in kcal mol^{-1} $kbar^{-2}$. NMR calculations were performed by T. Pongratz using the chemical shift correction from Table 28.[224] The fit takes the dominant conformer from direct calculations (P_{II} for Ac-Ala-NHMe and α_R for Ac-Gly-NHMe) and yields fit parameters for the free energy difference of the second most dominant conformer (α_R for Ac-Ala-NHMe and P_{II} for Ac-Gly-NHMe).

Molecule/Geometries	ΔG_0	*s*	*t*
Ac-Ala-NHMe/ PCM	0.368	-0.021	0.0012
Ac-Ala-NHMe/ EC-RISM	0.025	-0.206	0.063
Ac-Gly-NHMe/ PCM	-2.467	0.130	-0.854
Ac-Gly-NHMe/ EC-RISM	0.326	-0.079	0.015

Table 30: Overview of populations and free energies for the two most common conformers of Ac-Ala-NHMe and Ac-Gly-NHMe, respectively. Theoretical values from EC-RISM calculations with MP2/6-311+G(d,p) on B3LYP/6-311+G(d,p)/PCM- or B3LYP/6-311+G(d,p)/EC-RISM-optimized structures are given as a comparison, once using the direct calculation and the PMV integration according to eq. (105). The pressure-dependent correction of the excess chemical potential was used for the direct calculation . Results are taken from calculations performed by T- Pongratz[224] which for the NMR fit use the here developed correction scheme for the amide chemical shifts with fit parameters ΔG_0, *s* and *t* from Table 29.

			Opt: B3LYP/6-311+G(,p)/PCM							Opt: B3LYP/6-311+G(,p)/EC-RISM						
Ac-Ala-NHMe			EC-RISM$_{calc}$			PMV-int.		NMR fit		EC-RISM$_{calc}$			PMV-int.		NMR fit	
Conf	φ/ °.	ψ/ °	1 bar	1 kbar	5 kbar	1 kbar	5 kbar	1 bar	3 kbar	1 bar	1 kbar	5 kbar	1 kbar	5 kbar	1 bar	3 kbar
P_{II}	-76	143	58.95	58.46	57.81	58.64	57.47	65.04	63.07	64.75	65.81	65.27	64.40	63.27	51.04	48.74
α_R	-90	-12	32.76	33.60	34.7	33.18	34.57	34.96	36.93	27.34	26.81	27.77	27.78	29.12	48.96	51.26
			ΔG / kcal mol^{-1}													
P_{II}	-76	143	0	0	0	0	0	0	0	0	0	0	0	0	0	0
α_R	-90	-12	0.35	0.34	0.30	0.34	0.30	0.37	0.32	0.51	0.53	0.51	0.50	0.46	0.02	-0.03
Ac-Gly-NHMe																
Conf	φ/ °.	ψ/ °	1 bar	1 kbar	5 kbar	1 kbar	5 kbar	1 bar	3 kbar	1 bar	1 kbar	5 kbar	1 kbar	5 kbar	1 bar	3 kbar
α_R	-98	-1	31.4	31.57	31.49	31.43	31.34	1.53	0	29.83	29.72	30.5	29.91	30.04	36.59	40.78
P_{II}	78	-154	13.24	13.13	13.2	13.2	13.21	98.47	100	16.22	15.67	15.9	16.07	15.66	63.41	59.22
			ΔG / kcal mol^{-1}													
α_R	-98	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P_{II}	78	-154	0.49	0.49	0.49	0.49	0.48	-2.46	-8.74	0.36	0.38	0.39	0.37	0.39	-0.33	-0.22

Table 29 shows that the pressure-dependent free energy change to match experimental chemical shift changes is low with the only exception being Ac-Gly-NHMe using PCM geometries which seems to be an outlier and unreliable as it is the only case were one conformer (α) more closely matches both experimental chemical shifts even after the correction (see Figure 33 **G** and **H**). In the case where this happens the NMR fit fully shifts the conformational equilibrium to the α conformer even at ambient conditions as seen in Table 30. Using EC-RISM geometries this behavior changes (see Figure 29 **G** and **H**) leading to a more balanced NMR fit as moving the equilibrium to one conformer would benefit one chemical shift but worsen the other. For Ac-Ala-NHMe this is found for both sets of geometries resulting in small fit parameters. It is noteworthy that the fit parameters *s* and *t* show opposing signs in all cases reducing the overall effect of both. Without Ac-Gly-NHMe/PCM, which is treated as an outlier, the free energy difference between the two main conformers only changes by up to 0.11 kcal/mol at a pressure of 3 kbar as seen in Table 30, confirming direct calculations where both calculation methods agree with this overall very small change in the free energy.

This also matches one observation from the initial NMR calculations. The overall pressure dependence of the chemical shift is already well matched by each individual conformer (see Figure 33 or Figure 29 **D**, **E**, **I**, **J**), however each individual conformer has a distinct chemical shift differing by up to 0.3 ppm

for Ac-Gly-NHMe regardless of the fit. If the population changes drastically with an increase in pressure the resulting mean chemical shift would also change because of the chemical shift differences between conformers. In this way the absolute accuracy of chemical shift predictions becomes important which here can only be reached by fitting to training data.

The experimental pressure dependence of the chemical shift of the peptide amides already matches well with the other amides investigated as part of the amide subset, see Figure 25, for which no conformational changes are expected at high-pressure conditions. In this way only the intrinsic pressure-dependent changes in solvation and geometry (i.e. bond lengths) are seen for the two dipeptides investigated here.

Overall, this investigation is a successful proof of concept for fitting pressure-dependent NMR data as a second method to determine conformational equilibria without direct energetic calculations. The necessary accuracy can be reached by single structure calculations if a training dataset of comparable molecules is available. The benefit of this method is in the high accuracy of NMR measurements yielding reliable experimental data even for extreme conditions such as high hydrostatic pressure.

4.4.3 Summary and outlook

Within this chapter the pressure-dependence of the chemical shift was investigated using a combination of MD simulations (both AIMD and FFMD) and single point methods including newly developed EC-RISM geometry optimizations.

Here, the amide subset, measured within this thesis, was used as a benchmark dataset for theoretical predictions. For this subset EC-RISM geometries yielded the best overall results at ambient conditions with RMSEs of 0.76 ppm for ^{1}H, 2.37 ppm for ^{13}C and 5.02 ppm for ^{15}N. The use of EC-RISM geometries improves the accuracy of predictions beyond the FFMD-based structural ensembles, which, in chapter 4.2 were still on par with the then used PCM geometries. Together with its small computational demand compared to FFMD-based methods this clearly makes single point calculations on EC-RISM geometries the method of choice for ambient-condition NMR predictions.

The use of FFMD-based structural ensembles overall yields non satisfactory results at ambient conditions, in particular for aromatic protons, carbonyl-carbon and the ^{15}N reference ammonia, showing the importance of the specific force field that is used, and once again showing specific nuclei which need further improvement.

Here the use of MLMD- or AIMD simulations as shown before could also yield more accurate structural ensembles. Whenever these higher-level simulation methods are not feasible the use of EC-RISM geometries for single structure calculations is still recommended as classical force field might not be able reach the accuracy provided by EC-RISM and still require extensive sampling to reduce statistical errors.

Using PCM- or EC-RISM geometries could yield accurate results for the amide protons after fitting the data to experimental ones. Here the intrinsically larger error found for the PCM-geometries could also be removed by a linear fit and afterwards all amide chemical shifts could be matched to experiments with an error of less than 0.21 ppm.

In collaboration with T. Pongratz[224] the fitting parameters were then applied to two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe, which drastically reduced the errors found for the amide protons of these dipeptides. Using the corrected chemical shifts, the conformational equilibrium, undetectable by direct NMR measurements, could then be fitted to match the conformation-weighted experimentally determined pressure-dependent chemical shift changes. The population changes derived from this fit showed a small pressure-dependence of the conformational equilibrium only changing by 0.1 kcal/mol at 3 kbar, conforming previous direct calculations in a completely independent way. This approach allows for further similar investigations of conformational equilibria based on accurate chemical shift predictions and averaged NMR measurements combining spectroscopy with thermodynamics.

For the prediction of the pressure-dependence of the ^{13}C and ^{15}N chemical shifts of the amide subset only the fully solvated FFMD snapshots at least give the right overall trend with a positive slope of the linear regression. This, however, comes with the largest RMSE for the ^{15}N predictions revealing another important feature detected during this study. The accurate modeling of the reference compounds DSS and ammonia under high pressure turned out to be difficult task. For DSS experimental data shown in earlier chapters indicated a decrease in the shielding constants, which not only was not found previously[7] but during this investigation was only found with the FFMD-based methods.

Here the use of AIMD or even MLMD simulations could be the key to accurately modeling DSS also under high-pressure conditions. With an accurate reference, the relative trend of the ^{13}C chemical shift differences would already be well matched by the FFMD using fully solvated snapshots and absolute errors could be decreased.

In terms of the pressure-dependence of the ^{15}N chemical shifts the referencing method is still questionable. The benchmark data was created using a small pressure-dependence of the reference with a pressure-independent Ξ value. In contrast the fully solvated snapshots showed a large pressure-dependence of ammonia worsening the overall metrics. A further investigation of the ^{15}N chemical shift reference like done here in chapter 4.3 with the ^{13}C reference is necessary to create reliable ^{15}N data. This investigation could consist of the use of ammonia as a direct reference or the use of another reference compound such as nitromethane. As the theoretical methods heavily disagree on the ^{15}N chemical shift reference under pressure an accurate experimental reference is necessary but does not exist at this point.

With the methods used during this investigation, disentangling the structural and solvation response to high pressure was one of the main goals. The fact that only snapshots with explicit solvation instead of the "$FFMD_{VD}$" method yielded accurate pressure-dependent chemical shift trends together with the fact that using the pressure-dependent EC-RISM geometries only slightly increases prediction

performance showed that the response to high pressure is mainly driven by the change in solvation. This extends the findings from chapter 4.1.2 to a broader set of nuclei for which a similar behavior was found.

With more experimental data available for the whole high-pressure dataset further investigations are possible in the future to also investigate the set of amino acids that were measured at high pressure. This investigation could also show how different protonation states behave under high-pressure conditions. However, this would require simulations or single point calculations for all possible protonation states which were not feasible here.

4.5 pH dependency of the EPR spinprobe HMI

Extending on a previous set of publications[9][248] where calculations only on the neutral from of HMI were made, the goal of this study was to understand the pH sensitivity of the spin probe HMI. HMI is long known as a pH sensitive spin probe[47] and to this day shows one of the largest changes in A_{iso} upon protonation among nitroxide spin labels.[48] With that it is an ideal spin probe to detect changes in the local pH value, however the underlying structural reasons for this pH dependent change in EPR parameters and determining these reasons was the main goal of this study.

The previous publications[9][248] established a workflow of AIMD simulations followed by a hybrid solvation approach, similar to a “local,ECR” or “local,QM/MM” approach as described in chapter 4.2. With that high level CCSD calculations were able to quantitatively match experimental A_{iso} values,[9] while hybrid DFT calculations were able to accurately represent experimental trends. [248] Here it was also shown for HMI, that different populations with distinguishable EPR parameters exist and result in a strain of the *g*-tensor. These could be related to different hydrogen bond patterns around the nitroxide.

These approaches are now extended to the protonated form of HMI ($HMIH^+$) using experimental measurements, performed by L. Galazzo, AIMD simulations, performed by B. Sharma, and EPR calculations with a hybrid solvation approach.[267] As the main focus is on understanding the changes between HMI and $HMIH^+$ a hybrid DFT level of theory is deemed sufficient to accurately reproduce experimental trends and will be used throughout for all EPR calculations. By varying the description of the solvent the pH dependency will be split into a structural and solvation part, which will be quantitatively determined.

4.5.1 Comparison of HMI and $HMIH^+$

The comparison between HMI and $HMIH^+$ starts with aA comparison of spectra at pH 2 and pH 10 which is shown in Figure 33.

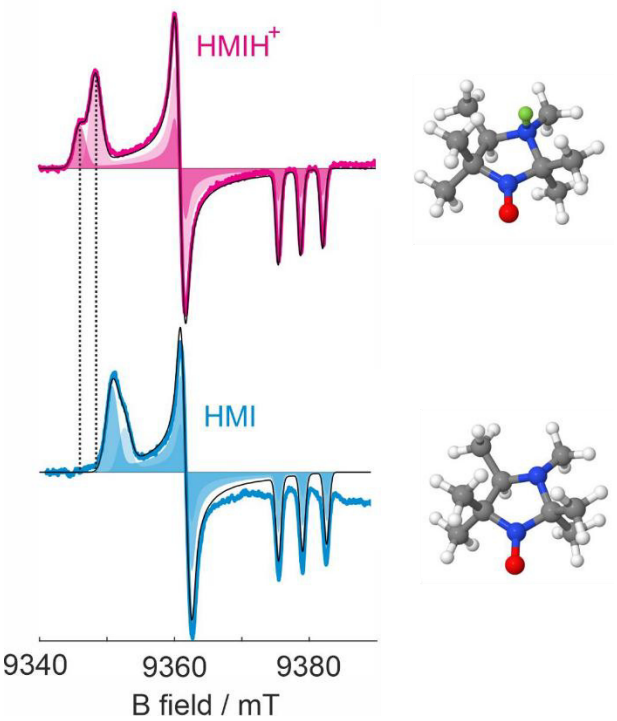

Figure 30: Experimental spectra of HMI (bottom, measured at a pH of 10) and $HMIH^+$ (top, measured at a pH of 2). The three-dimensional structure corresponding to (3R)-HMI and (3R,4S)-$HMIH^+$ are shown aswell. Spectra were recorded and figure was designed by L. Galazzo.[267]

Experimentally the following EPR-parameters were obtained: A^{iso} = 44.87 ± 0.14 MHz and g^{iso} = 2.00550 ± 10^{-5} for HMI at a pH value of 10, and A^{iso} = 41.20 ± 0.14 MHz and g^{iso} = 2.00575 ± 10^{-5} for $HMIH^+$ at a pH value of 2 showing a significant pH dependence for both observables. Again, for $HMIH^+$ individual subpopulations are visible, leading to the same assumption[248] that distinguishable hydrogen-bond subensembles are present for $HMIH^+$.

With the protonation of HMI, two diastereomers could potentially exist. The (3R,4S), shown in Figure 33, and the (3R,4R) with the additional proton on the other side. First calculations were performed to identify the relevance of both diastereomers. Using an indirect approach with solvation free energies on the revPBE0-D3/EC-RISM level of theory with DLPNO-CCSD(T) gas phase energies showed that the (3R,4S) configuration is favored by 3.54 kcal mol^{-1} (using the PMV correction taken from Ref.[175] and 3.08 kcal mol^{-1} without the PMV correction) compared to the (3R,4R) configuration. Hence only the (3R,4S) configuration was used in AIMD simulations.

These simulations, performed by B.Sharma,[267] showed a significant difference in the coordination number and an average number of H-bonds as shown in Table 31. At a higher pH value, the average number of H-bonds surrounding the nitroxy-oxygen increases from 1.58 for $HMIH^+$ to 2.16 for HMI. This can be attributed to an increase in snapshots containing three hydrogen bonds from 6.6% to

27.2%, changing the second most populated sub-ensemble from one H-bond for $HMIH^+$ to three H-bonds for HMI. This shows that even though the protonatable group is not directly part of the EPR active nitroxide the specific protonation pattern still heavily impacts the solvation of the nitroxide group.

Table 31: Comparison of H-bond statistics for HMI and $HMIH^+$ in water where CN, $\langle n_{HB} \rangle$, and P_n are coordination number, average number, and population with n H-bonds between the nitroxy-oxygen and the water-oxygens, respectively. The populations of the two major fractions are highlighted in bold. The simulations and H-bond analysis were performed by B. Sharma. The H-bond criterion $r(O\cdots H)<1.71\cos\vartheta+1.37$ as established in Ref.[9] for HMI was also used for $HMIH^+$.

Property	$HMIH^+$	HMI
CN	2.70	2.80
$\langle n_{HB} \rangle$	1.58	2.16
P_0	4.4%	0.3%
P_1	**40.0%**	12.0%
P_2	**49.0%**	**59.8%**
P_3	6.6%	**27.2%**
$P_{>3}$	0.0%	0.7%

Based on this finding the quantitative influence of the change in coordination was further investigated. By varying the used solvation model structural and solvent effects on EPR parameters could be disentangled. Therefore, a hierarchy of solvation models analogous to the NMR calculations was used. First a vertically desolvated ensemble (VD) where the solute was stripped of all solvent during the EPR calculation to establish a baseline only containing solvation effects from the AIMD simulations implicitly based on difference in solute structures. Additionally an explicit second solvation shell of water molecules around the nitroxy oxygen (SSS) was added and two hybrid solvation models were used by using either EC-RISM or force field-based background charges (QM/MM) on top of SSS snapshots. Compared to the NMR calculation this follows the “local” explicit solvation approach with a radius of 5 Å around the nucleus of interest. All models containing explicit solvation are further analyzed on the level of individual sub-ensembles defined by the number of H-bonds as these show a significant change in population upon protonation as shown before in Table 31.

The results for the *g*- and *A*-tensor and their components are summarized in Table 32 for experimental results and using the QM/MM and EC-RISM solvation methods. It is noteworthy that individual tensor components were experimentally investigated at 100 K. This results in a slightly different A^{iso} as the trace of the tensor components compared to the values reported in Table 33. As the AIMD simulations were also performed at ambient conditions however comparing them to ambient condition X-band

results is a more consistent approach and for the comparison to low temperature J-band measurements trends should be comparable while absolute values can already differ based on the different conditions. Earlier investigations on the temperature dependence[268][269][270] showed an increase in A^{iso} with increasing temperature for 5-membered ring nitroxides.

Table 32: Parameters used to fit the components ("Comp1", "Comp2") of the experimental spectra of $HMIH^+$ and HMI compared to the theoretical parameters ("TComp1", "TComp2"). Experimental data was measured by L. Galazzo. Populations (weights) for QM/MM and EC-RISM were determined by B.Sharma from AIMD simulations. To better compare with the experimental spectra simulated with two components, weights for only the two major theoretical populations were renormalized. These (namely one and two H-bonds) contribute to 89% of the cases for HMIH+, and the two and three H-bonds sub-ensembles for HMI, that contribute to 87% of the cases. Raw data is published in the electronic supplementary material, see chapter 6.1.

		experimental		**QM/MM**		**EC-RISM**	
		Comp1	*Comp2*	*TComp1*	*TComp2*	*TComp1*	*TComp2*
$HMIH^+$	# H-bonds			1	2	1	2
	weights	0.33	0.67	0.45	0.55	0.45	0.55
	g_{xx}	2.00937	2.00887	2.00855	2.00829	2.00837	2.00817
	$(g_{xx}-g_{zz})/10^{-5}$	700	650	643	618	625	605
	g_{yy}	2.00617		2.00585	2.00581	2.00582	2.00577
	g_{zz}	2.00237		2.00212	2.00211	2.00212	2.00212
	A_{xx} [MHz]	13.0		6.0	6.6	6.4	7.0
	A_{yy} [MHz]	13.0		6.4	6.9	6.7	7.2
	A_{zz} [MHz]	92.2		85.2	88.9	87.9	90.8
HMI	# H-bonds			2	3	2	3
	weights	0.67	0.33	0.69	0.31	0.69	0.31
	g_{xx}	2.00834	2.00795	2.00787	2.00763	2.00778	2.0760
	$(g_{xx}-g_{zz})/10^{-5}$	604	565	575	552	586	549
	g_{yy}	2.00598		2.00572	2.00568	2.00570	2.00567
	g_{zz}	2.00230		2.00212	2.00211	2.00212	2.00211
	A_{xx} [MHz]	14.0		7.3	7.4	7.5	7.8
	A_{yy} [MHz]	14.0		7.6	7.7	7.4	7.8
	A_{zz} [MHz]	100.0		95.3	98.2	96.5	98.7

Table 32 shows that the populations of the two main ensembles agree well with experimentally determined populations also for $HMIH^+$ which validates the hydrogen-bond based assignment of subensembles performed in this case. Even though the EPR calculations were performed on the DFT level (revPBE0-D3 functional with the def2-TZVPP basis set with decontracted s-functions) and it is well-known that quantitative agreement can only be approached with coupled cluster theory[9] these DFT calculations still agree with experiment on the trends that can be found. The difference g_{xx}-g_{zz} shows a difference between the two main ensembles of $50 \cdot 10^{-5}$ ($HMIH^+$) and $39 \cdot 10^{-5}$ (HMI). EC-RISM here yields $20 \cdot 10^{-5}$ and $37 \cdot 10^{-5}$ respectively and QM/MM yields $25 \cdot 10^{-5}$ and $23 \cdot 10^{-5}$ respectively and individual g_{xx}-g_{zz} for each subensemble only differs by up to 10%.

With all tensor components also known from theoretical calculations it is possible to simulate the resulting J-band spectra and compare with experimentally measured spectra. This is shown in Figure 31.

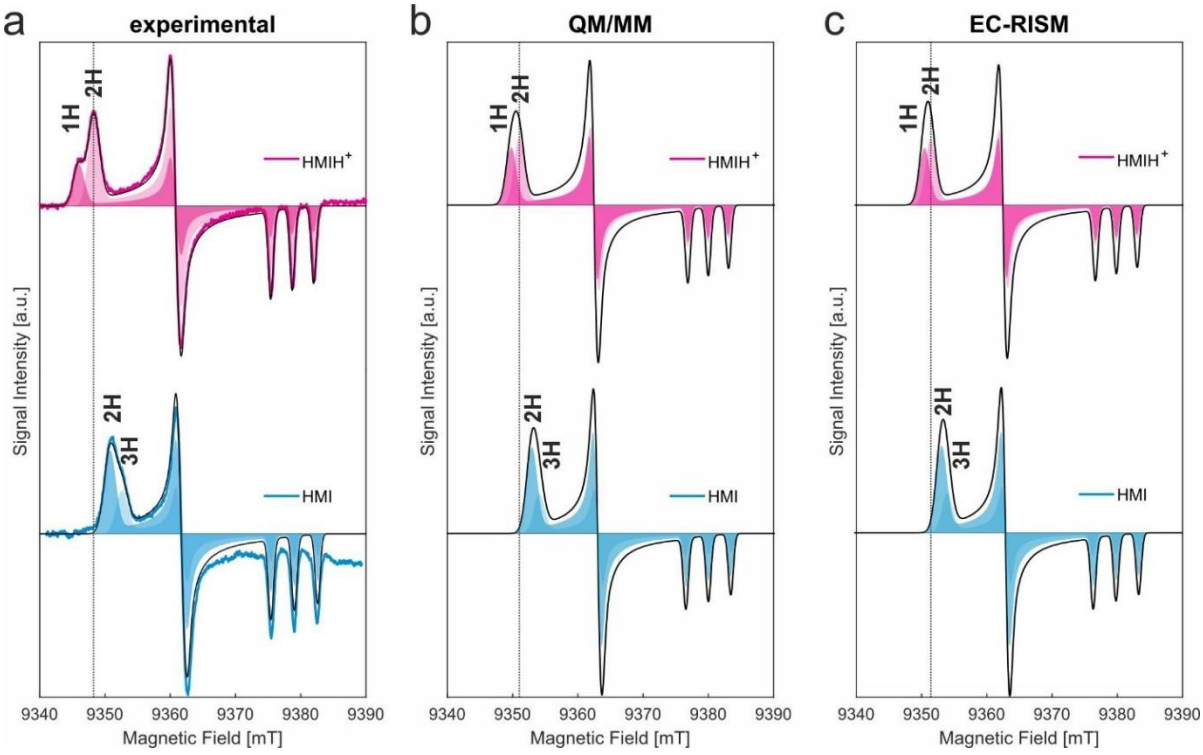

Figure 31: EasySpin simulations of the J-band EPR spectra of HMI and HMIH+ obtained using the parameters shown in Table 32 for the experimental parameters and based on EC-RISM- and QM/MM calculations. The two spectral components considered in the simulations are highlighted as shaded areas of different transparency, respectively. The weighted sum spectrum is shown in black. Recorded experimental spectra (in magenta and blue) are overlaid to the simulated spectra in the left panel. A grey vertical line was added at the magnetic field corresponding to the highest g value per method (experimental, QM/MM, EC-RISM) for HMIH+ to highlight differences between protonation states. The simulation of spectra was performed by L. Galazzo.[267]

Figure 31 shows that on the level of spectra theory and experiment visibly match almost perfectly. The shift to higher magnetic field (lower *g* value) for unprotonated HMI as the most prominent feature can be seen for QM/MM and EC-RISM as well. The hyperfine coupling most prominent for A_{zz} also shows good agreement for QM/MM and EC-RISM. From Table 32 the error in A_{zz} can be quantified to be less than 7 MHz ($HMIH^+$, QM/MM, comp. 1) which still yields very similar spectra as shown in Figure 31.

In a liquid solution at ambient temperature, the EPR spectra, as shown in Figure 33, will only reflect the H-bond population-weighted average of the *g* and *A* tensorial quantities of HMI and $HMIH^+$. Therefore, the isotropic quantities A^{iso} and g^{iso} of the complete ensemble can first be compared with the quantities calculated for vertically desolvated HMI and $HMIH^+$. In a second step the relative protonation effect based on the differences between $HMIH^+$ and HMI has to be viewed method-specifically from the angle of individual H-bond contributions, which were just shown to be significant and validated based on comparison to J-band EPR spectra.

All results are summarized in Table 33 including statistical uncertainties based on error propagation assuming a normal distribution.

Comparing absolute values for the whole ensemble using vertically desolvated snapshots it becomes obvious that solvation during the EPR calculation contributes majorly to the absolute value of EPR parameters as seen before for NMR parameters. Here the explicit second solvation shell (SSS) already captures the majority of the solvation effect on A^{iso} and g^{iso} (see lines "[SSS|QM/MM|EC-RISM]-VD"). Further polarization from background charges adds a smaller increment which is more relevant for $HMIH^+$ than for HMI, from 4.9 (SSS) to 6.1 MHz (EC-RISM) for A^{iso} of HMI and from 4.0 to 6.6 MHz for A^{iso} of $HMIH^+$. In contrast, g^{iso} appears to be less influenced by background polarization beyond the SSS, indicating a more locally confined solvation effect. This is also corroborated by closer correspondence between QM/MM and EC-RISM for g^{iso}. This is similar to the NMR calculations where the explicit solvation also accounted for most of the solvation effect. Once again EC-RISM seems more accurate than QM/MM in terms of background solvation as it captures more of a solvation effect, this was also visible in the convergence study for NMR parameters.

Analyzing the individual ensembles based on the number of H-bonds two major points become clear. Regardless of the specific solvation method both A^{iso} and g^{iso} change with an increasing number of H-bonds. Here A^{iso} continuously increases with an increase in the number of H-bonds while g^{iso} decreases. Strikingly this effect is much smaller for the difference between HMI and $HMIH^+$ (see columns ΔA^{iso} and Δg^{iso}) indicating that a consistent ΔA^{iso} and Δg^{iso} if the solvent surrounding the nitroxy group is fixed and only the protonation changes.

Additionally, the relative protonation effect based on the differences between $HMIH^+$ and HMI has to be viewed method-specifically from the angle of individual H-bond contributions. Even though the electronic structure effect captured by the VD ensemble seemingly accounts for the total relative change of both A^{iso} and g^{iso} upon protonation, including solvation by using either the SSS, EC-RISM, or QM/MM approaches clearly show strong solvation effects on absolute numbers. This effect, as previously mentioned, is larger for $HMIH^+$ compared to HMI which also impacts ΔA^{iso} and Δg^{iso} that

therefore are different for each solvation method. Further dissection by analyzing individual H-bond sub-ensembles will provide quantitative insight into the solvation effect on ΔA^{iso} and Δg^{iso} since, as shown in Table 31, their populations change strongly upon protonation.

A compact summary of individual H-bonded populations (in terms of one, two or three H-bonds) on A^{iso} and g^{iso} values and their differences between protonated and neutral forms is provided in Table 33 When comparing the individual H-bond ensembles to the overall ΔA^{iso} and Δg^{iso} for each method (SSS|QM/MM|EC-RISM) it is noteworthy that the overall ΔA^{iso} and Δg^{iso} (see the bold lines SSS|QM/MM|EC-RISM) is by magnitude always larger than for any individual H-bond ensemble using the same method. This can only be the result of the change in H-bond populations and allows to quantify this effect. By taking the average of the individual subensembles (denoted by $\langle\Delta$H-bonds$\rangle_{SSS|EC\text{-}RISM|QM/MM}$) the resulting ΔA^{iso} and Δg^{iso} show a hypothetical situation where H-bond populations would not change. For example, individual H-bonding states yield a difference in A^{iso} of 2.2±0.2 MHz for EC-RISM (denoted by $\langle\Delta$H-bonds$\rangle_{EC\text{-}RISM}$) and 2.7±0.2 MHz for QM/MM ($\langle\Delta$H-bonds$\rangle_{QM/MM}$). If the population of H-bond ensembles remained constant between HMI and $HMIH^+$ these values would match the total ΔA^{iso} and Δg^{iso} change for EC-RISM or QM/MM. However, this is not the case as the average over all H-bond sub-ensembles for all methods and both parameters, ΔA^{iso} and Δg^{iso}, consistently yields only ca. 80% of the total differences (denoted by $\langle\Delta$H-bonds$\rangle_{SSS|EC\text{-}RISM|QM/MM}/\Delta$(SSS|EC-RISM|QM/MM). The remaining 20% can then directly be attributed to the change in H-bond population allowing the quantitative dissection of solvent effects from structural effects.

Table 33: A^{iso} and g^{iso} values including statistical uncertainties of HMI and $HMIH^+$ for VD, SSS, EC-RISM, and QM/MM calculations: Population-weighted averages over all H-bond sub-ensembles (bold), difference to VD results, population-independent individual values per sub-ensemble, and (for ΔA^{iso} and Δg^{iso} only) population-independent averages over all H-bond sub-ensembles. A graphical illustration of the EC-RISM and QM/MM parameters for each H-bond sub-ensemble is shown in Fig. 4. For ΔA^{iso} and Δg^{iso} the exact values were taken for computing the difference and rounded afterwards. Statistical uncertainties were calculated as standard errors under the assumption of a normal distribution and statistically uncorrelated snapshots. Uncertainties of differences were estimated by error propagation. Raw data is published in the electronic supplementary material, see chapter 6.1.

	HMI		$HMIH^+$		HMI-$HMIH^+$	
Method	A^{iso} [MHz]	g^{iso}	A^{iso} [MHz]	g^{iso}	ΔA^{iso} [MHz]	Δg^{iso}
Experimental$^{298\ K}$	**44.9**	**2.00550**	**41.2**	**2.00575**	**3.7**	**-0.00025**
VD	**31.2(2)**	**2.00574(2)**	**27.8(2)**	**2.00589(1)**	**3.4(2)**	**-0.00015(2)**
SSS	**36.1(2)**	**2.00527(1)**	**31.8(2)**	**2.00554(1)**	**4.3(2)**	**-0.00027(1)**
SSS-VD	4.9(2)	-0.00047(2)	4.0(2)	-0.00035(1)	0.9(3)	-0.00016(2)
SSS, 1 H-bond	34.4(5)	2.00539(1)	31.0(3)	2.00559(1)	3.3(5)	-0.00020(2)
SSS, 2 H-bonds	36.0(2)	2.00528(1)	32.5(3)	2.00550(1)	3.5(3)	-0.00022(1)
SSS, 3 H-bonds	37.1(3)	2.00519(1)	33.2(6)	2.00544(2)	3.9(7)	-0.00025(2)
$\langle\Delta$H-bonds$\rangle_{SSS}$	-	-	-	-	3.6(3)	-0.00022(1)
$\langle\Delta$H-bonds$\rangle_{SSS}$/ ΔSSS	-	-	-	-	83%	81%
QM/MM	**36.8(2)**	**2.00523(1)**	**33.4(2)**	**2.00546(1)**	**3.4(2)**	**-0.00023(1)**
QM/MM-VD	5.6(2)	-0.00052(2)	5.6(2)	-0.00043(2)	0.0(3)	-0.00008(2)
QM/MM, 1 H-bond	34.9(5)	2.00536(1)	32.5(3)	2.00551(1)	2.3(5)	-0.00015(2)
QM/MM, 2 H-bonds	36.8(2)	2.00524(1)	34.1(3)	2.00541(1)	2.6(3)	-0.00017(1)
QM/MM, 3 H-bonds	37.8(3)	2.00514(1)	34.7(6)	2.00534(2)	3.0(6)	-0.00020(2)
$\langle\Delta$H-bonds$\rangle_{QM/MM}$	-	-	-	-	2.7(3)	-0.00017(1)
$\langle\Delta$H-bonds$\rangle_{QM/MM}$/ ΔQM/MM	-	-	-	-	78%	75%
EC-RISM	**37.3(2)**	**2.00520(1)**	**34.4(2)**	**2.00540(1)**	**2.9(2)**	**-0.00020(1)**
EC-RISM-VD	6.0(2)	-0.00055(2)	6.6(2)	-0.00049(2)	-0.5(3)	-0.00005(2)
EC-RISM, 1 H-bond	35.6(5)	2.00531(1)	33.7(3)	2.00544(1)	2.0(5)	-0.00013(2)
EC-RISM, 2 H-bonds	37.3(2)	2.00520(1)	35.0(3)	2.00536(1)	2.3(3)	-0.00015(1)
EC-RISM, 3 H-bonds	38.0(3)	2.00513(1)	35.6(6)	2.00529(2)	2.4(6)	-0.00016(2)
$\langle\Delta$H-bonds$\rangle_{EC\text{-}RISM}$	-	-	-	-	2.2(3)	-0.00015(1)
$\langle\Delta$H-bonds$\rangle_{EC\text{-}RISM}$/ ΔEC-RISM	-	-	-	-	77%	75%

A compact illustration of these findings Is given in Figure 32 showing the increase in A^{iso} and decrease in g^{iso} with an increasing number of H-bonds while ΔA^{iso} and Δg^{iso} stay more or less constant for each method. With the width of the bar the change in H-bond population can also be seen as the second most dominant ensemble has one H-bond for $HMIH^+$ and three H-bonds for HMI.

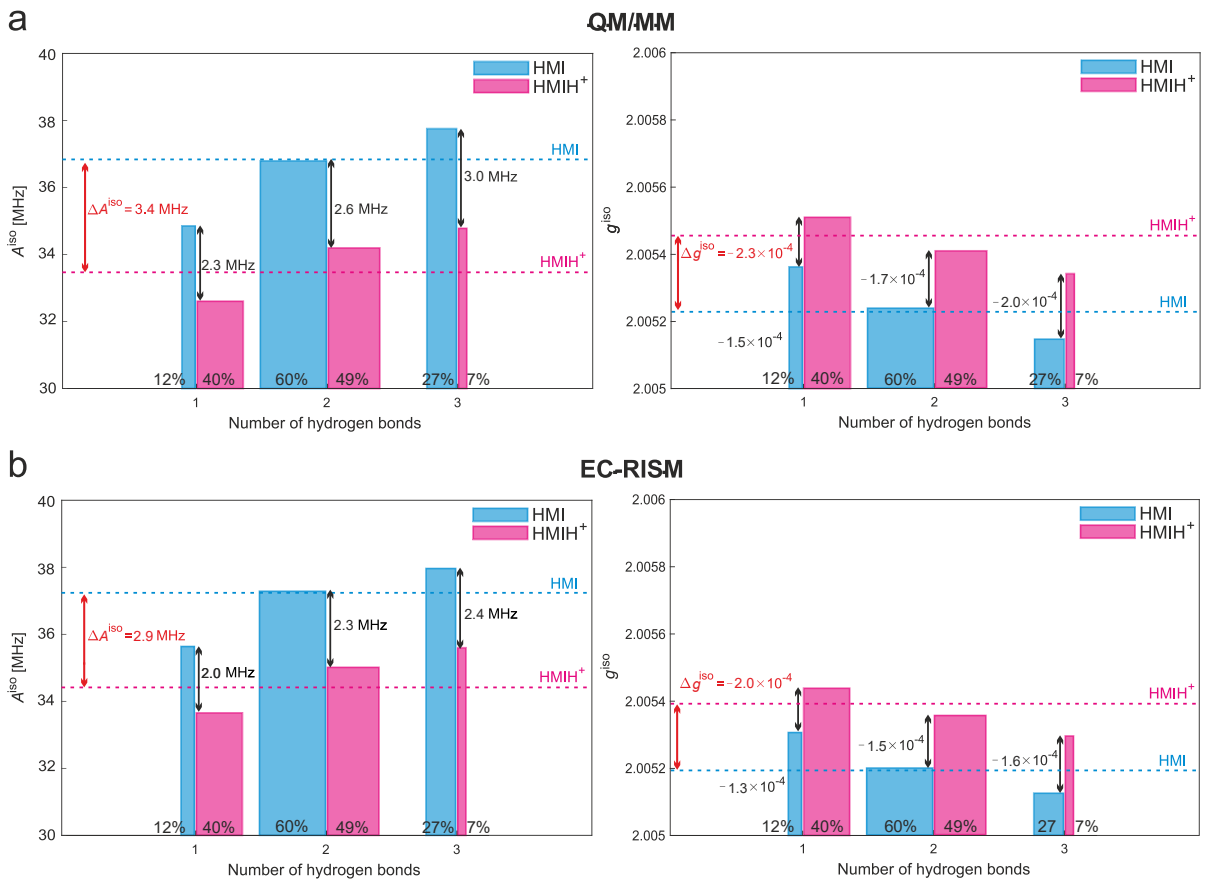

Figure 32: Computed A^{iso} (left panels) and g^{iso} (right panels) EPR parameters for HMI (blue) and $HMIH^+$ (magenta) with a fixed number of H-bonds using QM/MM (a) and EC-RISM (b). The width of each bar corresponds to the population of the specific sub-ensemble annotated by the number of H-bonds. Horizontal dashed lines show the final average values. Differences between sub-ensembles are shown with black arrows while the difference between the full ensembles is shown with a red arrow, where the solvation contribution is highlighted by a dashed line. A small number of structures (< 5%) corresponding to zero and more than three H-bonds was omitted for clarity as the statistical uncertainty was too high in these cases to gain insight. The figure was designed by L. Galazzo.

To summarize the finding: As the ca. 80% contribution per H-bond to the total average is consistent among all methods, and is found for both ΔA^{iso} and Δg^{iso}, this fraction is attributed to an electronic structure effect exhibited at a fixed solvent surrounding determined by the number of H-bonds and the specific method of background solvation. The remaining 20% can directly be assigned to the

change in population between H-bond sub-ensembles, the solvation effect of protonated versus neutral HMI in water.

4.5.2 Comparison of QM/MM and EC-RISM background solvation

As QM/MM and EC-RISM agree on all the trends shown within this chapter but disagree on specific values of i.e. in A^{iso} or ΔA^{iso} futher investigations into the underlying physical reasons have been made. The main differences between these methods are that EC-RISM uses a much larger solvent box compared to the spherical cutout of AIMD simulation boxes that are the basis of the QM/MM calculation. The second main difference is the use of discrete point charges in the QM/MM calculation while using EC-RISM the solvent is represented by continuous solvent distribution. To disentangle these two effects EC-RISM calculations have been performed using a smaller box comparable to the size of the QM/MM point charges. The difference between the EC-RISM calculation for the large box and the smaller box can be attributed to the effect of the box size and the remaining difference compared to the QM/MM values can be attributed to the discretization of point charges. As EC-RISM calculations with this small box size ($46^3 \cdot 0.5$ Å) are difficult to converge only a limited number of snapshots were used. The overview of A^{iso} and g^{iso} for all methods per snapshot is given in Table 34.

Table 34: A^{iso} and g^{iso} for a reduced subset of nine snapshots ($HMIH^+$) and 14 snapshots (HMI) which have converged using EC-RISM with a small box size ($46^3 \cdot 0.5$ Å) comparable to the QM/MM box size. Raw data is published in the electronic supplementary material, see chapter 6.1.

	HMI		$HMIH^+$	
Method	A^{iso} [MHz]	g^{iso}	A^{iso} [MHz]	g^{iso}
EC-RISM	36.2	2.00510	32.3	2.00528
QM/MM	35.8	2.00513	31.4	2.00533
SSS	35.5	2.00515	29.5	2.00545
$EC\text{-}RISM_{small\ box}$	35.9	2.00512	31.7	2.00532

Table 34 shows that the majority of the difference from QM/MM to EC-RISM can be attributed to the smaller range of point charges as both A^{iso} and g^{iso} are close to the QM/MM value for EC-RISM using the small box size. As a consequence the difference from using the discrete point charges compared to continuous solvent distribution is only marginal. The effect is also more pronounced for $HMIH^+$ compared to HMI as here the solvation of the protonated group of HMI seems to be important and using a small box size leads to larger errors. Even though the box size of 23^3 Å is already enlarged compared to the original AIMD box but still using a QM/MM solvation method based on this box

introduces errors especially in terms of absolute values of A^{iso} and EC-RISM is a preferred method for solvation even though it comes with a larger computational demand.

4.5.3 Summary and outlook

By extensively varying the solvent models used during EPR calculations the origin of the pH dependence of the EPR spin probe HMI was disentangled quantitatively. Here a hierarchy of solvent models was used. Starting from vertically desolvated (VD) snapshots from AIMD simulations an explicit second solvation shell was included resembling the "local" solvation approach outlined in the previous chapter. The addition of additional point charges (QM/MM) or EC-RISM background solvation was deemed sufficient in accurately describing the solvation around the EPR-active nitroxide, once again similar to the previously established for NMR calculations and earlier work on HMI.[9]

Even though AIMD simulations indicated a clear change in the hydrogen bonding populations around the EPR-active nitroxide group the main reason for the pH-dependent change in EPR parameters was found in the change in electronic structure upon protonation. Even though both A^{iso} and g^{iso} strongly depend on the number of hydrogen bonds, similar to the behavior found for chemical shifts in the previous chapter, the change in H-bond populations was found to only contribute 20% of the change of both A^{iso} and g^{iso}. The remaining ca. 80% of the change of both A^{iso} and g^{iso} can be attributed to the change in electronic structure. This was determined from a near constant ΔA^{iso} and Δg^{iso} for a fixed number of hydrogen bonds using a given background solvation. Both ΔA^{iso} and Δg^{iso} determined this way are also matched well by the VD approach where no solvation is present during the EPR calculation.

Compared to experiments absolute values of A^{iso} differ significantly, however this was also previously seen and is due to the DFT level of theory used in this work. High level CCSD(T) calculations have previously increased the absolute accuracy[9] however the use together with an explicit second solvation shell is not feasible at this point and the main goal of this project was to identify differences between the protonation states of HMI which was successfully done using the DFT level of theory.

Further comparing theoretical methods EC-RISM background solvation seemed to be more accurate compared to QM/MM. As part of a deeper investigation a small EC-RISM box comparable to the QM/MM zone was used for which results closely matched the QM/MM approach. This shows that long range interactions as present with standard EC-RISM calculations are necessary to be included. On the other hand, the discretization into point charges in the QM/MM approach does not show a large impact on resulting EPR parameters.

Based on these findings the use of HMI as a spin label can further be rationalized and more experimental investigations can be designed accordingly. When HMI is used to investigate different protein environments it is noteworthy that first and foremost a different electronic structure is

measured. On the one hand this can be due to a change in local pH, on the other hand other protein interactions can influence the electronic structure of HMI which changes the measured EPR parameters. Only secondarily a change in local solvation influences the EPR parameters with a ΔA^{iso} of ca. 0.7 MHz and a Δg^{iso} of ca. $5 \cdot 10^{-6}$ which can be attributed to the change in H-bond populations due to the change in solvation. Both values however are in the experimental range which can be detected by experiment so that just the change in solvation could also be detected.

From a theoretical point of view the workflow established here could also be transferred to other spin labels. While it is likely that the pH-dependence of other nitroxide spin probes is predominantly due to the change in the electronic structure systematic studies of this behavior would be necessary to confirm this. The workflow then contains of an AIMD simulation from which subensembles based on the H-bond pattern are taken and EPR calculations are performed using a hybrid explicit solvation together with preferably EC-RISM background solvation. This workflow should be able to also determine the origin of the pH dependence for other spin labels. It combines the high level structure generation using AIMD simulations together with a sufficient solvation model for the EPR calculation. By investigating a fixed solvent situation determined by a specific number of H-bonds the underlying reasons can be disentangled. This approach should be fully transferable to other EPR spin labels including but not limited to nitroxides.

4.6 Development of a solvated ANI-type Machine Learning Potential

The goal of this study together with the group of A. E. Roitberg is to develop a solvated ANI-type MLP based on aqueous EC-RISM calculations. With that the aim is to evaluate the possibilities of using this MLP in the standard workflow of property predictions, for which the MNSOL database, consisting of solvation free energies, and the Tautobase, consisting of tautomeric equilibria in water, are used as two benchmark dataasets. The thermodynamic cycle illustrating two different method for the calculation of tautomeric equilibria is given in Figure 33.

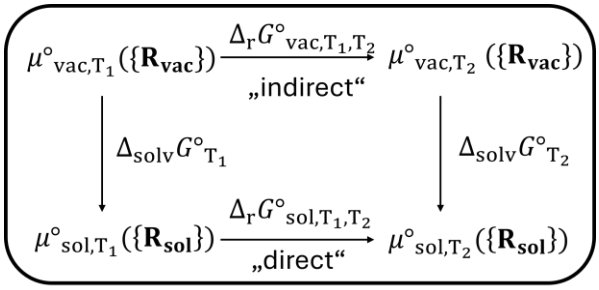

Figure 33: Schematic overview for two different calculation routes to calculate tautomeric equilibria. For direct calculation the difference of the free energy in solution is taken, while the indirect route makes use of the gas phase to calculate the free energy difference between tautomers.

Overall, three different MLPs are developed based on different observables of an EC-RISM calculation and the capability of predicting different properties shown in Figure 33 which will be outlined in more detail. These are then evaluated in terms of their direct property predictions and the application to geometry optimizations, a crucial step in all property predictions. This evaluation is done by first comparing the property prediction to direct high-level QM calculations using both MP2 calculations as well as the reference level of theory on which the MLP is created, r^{2}SCAN-3c, on the same set of geometries. Then the impact of geometry optimizations is evaluated by using the MLP as an input for MP2 calculations as well as MLP predictions themselves.

With this study an existing gap in the current literature is filled. Current general purpose MLPs typically are trained on gas-phase calculations.[66][271] The inclusion of solvent is then only realized by explicitly including solvent molecules. One prominent example of this are the MACE MLPs which were trained on ca. 10 % of the ANI-1x dataset and were close in terms of density and diffusion to AIMD simulations of liquid water.[272] The transferability of the MACE architecture to small organic molecules was also shown[67] outperforming ANI-type MLPs in terms of accuracy but at a higher computational cost due to the architecture of the MLP. In contrast the implicit solvation approach used here was only rarely followed with datasets like the AQM dataset[68] being smaller than the one developed here.

4.6.1 Overview of the developed Machine Learning Potentials

As the goal of this project was to develop a general purpose MLP a broad range of molecular structures and a large amount of training data is necessary. The number of calculations for each number of atoms for the ANI-EC-RISM dataset is shown in Figure 34.

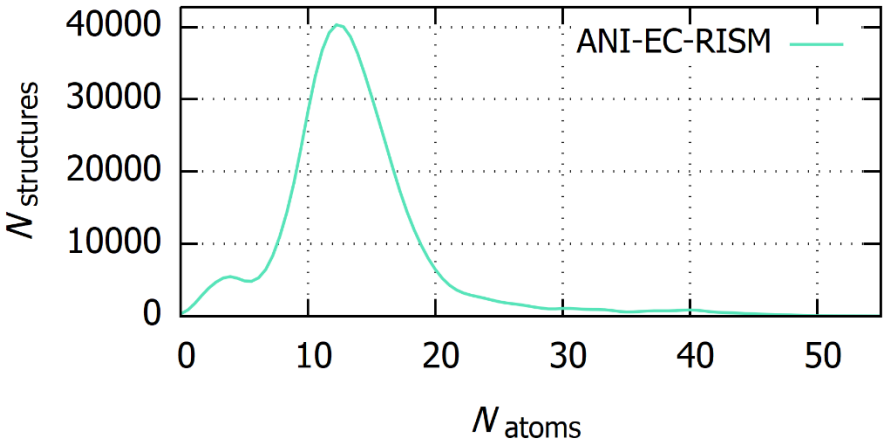

Figure 34: Kernel density estimate (KDE) of the number of structures of the ANI-EC-RISM dataset with respect to the number of atoms. The size of total number of structures is ca. 380000.

Figure 34 shows the composition of the ANI-EC-RISM dataset which, similar to ANI-2x (Figure 3) consists of small molecules with the most datapoints for molecules with ca. 9 to 14 atoms, including hydrogen. However the overall number of datapoints is only ca. 4 % of the original ANI-2x dataset. The ANI-EC-RISM dataset contains data of the coordinates and elements (species) of the molecule, taken directly from ANI-2x, and the resulting data from EC-RISM calculations shown in Table 35.

Table 35: Data with units contained in the ANI-EC-RISM dataset. GSAEs for G^{sol} and $G^{sol,corr}$ are taken from Table 5. $G^{sol,corr}$ was calculated with a C_V of of -0.09773 kcal/mol/Å^3, [273] corresponding to κ=0.714088·10^9 Pa^{-1}. All datasets are published in the electronic supplementary material, see chapter 6.1.

Data	Units
Coordinates	Å
Species	-
G^{sol}	E_h
$G^{sol,corr}$	E_h
V_m	Å^3
μ^{ex}	E_h

Individual MLPs were trained to the individual energetic outputs of an EC-RISM calculation and used accordingly in the following benchmark tests described in the next chapters. The validation RMSE values for each model are given in Table 36. The ANI-EC-RISM model is trained to G^{sol} while ANI-EC-RISM^{corr} already contains a PMV correction with C_v of -0.09773 kcal/mol/Å^3, corresponding to κ=0.714088·10^9 Pa^{-1}.[273] ANI-μ^{ex} is only trained to the μ^{ex}. Due to the low energetic range of the μ^{ex} no GSAEs were used in the development of ANI-μ^{ex}.

Table 36: Root mean squared error (RMSE) of the validation molecules in kcal/mol for the ANI models developed. The RMSE is based on a 80/20 training/validation split of the total dataset. All calculations were performed on the r^{2}SCAN-3c level of theory using the EC-RISM solvation model in water.

Model	RMSE / kcal/mol
ANI-EC-RISM	2.88
ANI-EC-RISM^{corr}	2.92
ANI-μ^{ex}	1.56

Table 36 shows that the validation RMSE is much smaller for ANI-μ^{ex} compared to the other models. This is because the range of data values is much smaller compared to ANI-EC-RISM and ANI-EC-RISM^{corr}. With errors of ca. 3 kcal/mol these two models still have a significantly larger intrinsic error compared to the original ANI-2x[66] (1.8 kcal/mol), however they only used ca. 4% of the training data. It is expected that a larger dataset also improves the validation RMSE getting close to the ANI-2x validation RMSE if the whole ANI-2x dataset is used. For example, during the development of this dataset a smaller dataset of ANI-EC-RISM containing 250 k datapoints had a validation RMSE of 3.1 kcal/mol.

The transferability of the training is first evaluated for all molecules contained in the Tautobase. Here a direct comparison between the training level of r^{2}SCAN-3c/EC-RISM and the MLP was performed on the same set of geometries. The result is shown in Figure 35.

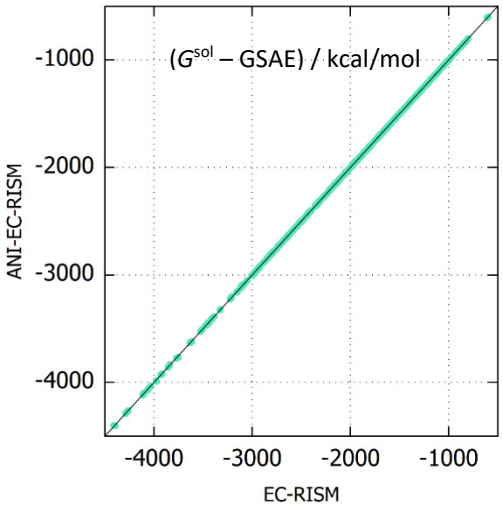

Figure 35: G^{sol} – GSAE in kcal/mol for the 910 molecules (including tautomers) of the Tautobase[274] for ANI-EC-RISM and EC-RISM using the r^2SCAN-3c level of theory. EC-RISM outputs were normalized using the GSAEs from Table 5 to match the direct output of the ANI-EC-RISM MLP. The RMSE between ANI-EC-RISM and EC-RISM is 2.46 kcal/mol. Both methods used the same set of geometries optimized at the B3LYP/6-311+G**/PCM level of theory provided by M. Strobl. Raw data is published in the electronic supplementary material, see chapter 6.1.

Figure 35 shows that even over a large range of G^{sol} – GSAE of more than 3000 kcal/mol the ANI-EC-RISM MLP can match direct EC-RISM calculations with an error of 2.46 kcal/mol. This is even lower than the validation RMSE of 2.88 kcal/mol during training showing that the ANI-EC-RISM is not overfitted to only predict the training dataset but can be transferred to an independent dataset. The fact that the test RMSE is even lower than the validation RMSE might be due to the fact that the training data is based purely on optimized structures while the whole dataset contains structures far from equilibrium. It is also possible that molecules from the Tautobase already occur in the ANI-2x dataset, however a detailed analysis of this was not performed and the Tautobase was treated as an independent test dataset as a whole.

4.6.2 Benchmark of a solvated ANI-type Machine Learning Potential on the MNSOL dataset

As a first benchmark dataset the Minnesota solvation database (MNSOL)[196] was used to benchmark the solvation free energy. As all ANI-EC-RISM potentials were only trained on neutral molecules only solvation free energies of neutral molecules in water were used leading to 374 data points. From this a further subset was created where no formally charged atoms, i.e. zwitterionic species were present. This resulted in 320 molecules which will be analyzed with ANI-EC-RISM.

In the prediction of solvation free energies two different methods were used.

First the direct the prediction of $G^{sol,corr}$ and subtracting gas phase energies calculated on the same level of theory according to equation (64). To accomplish this ANI-EC-RISM^{corr} was used to directly predict $G^{sol,corr}$ as a PMV correction was directly incorporated into the training data using a C_v of -0.09773 kcal/mol/Å^3, corresponding to κ=0.714088·10^9 Pa^{-1}.[273] With this, ANI-EC-RISM^{corr}, in theory, needs no further additional adjustment.

Second, the prediction of $\mu^{ex,corr}$ according to equation (65). With that $\mu^{ex,corr}$ would directly yield the solvation free energy if the molecular ensembles are the same in gas phase and in solution. To accomplish this the developed ANI-μ^{ex} needs a PMV correction which is performed here on existing data of MP2 calculations, provided by M.Strobl,[275] and also used to judge the reliability of this model. However, a successful development of ANI-μ^{ex} would not require the use of gas phase calculations in the prediction of solvation free energies, giving benefits to both approaches.

To benchmark the MLP both ANI-EC-RISM^{corr} as well as ANI-μ^{ex} were used while a PMV correction of ANI-μ^{ex} was fitted afterwards (with PMV data from MP2-calculations provided by M.Strobl[275]) The resulting data is comprised in Figure 36. Prediction metrics (RMSE, MSE and MAE) and regression parameters (m, b and R^2) are given in Table 37.

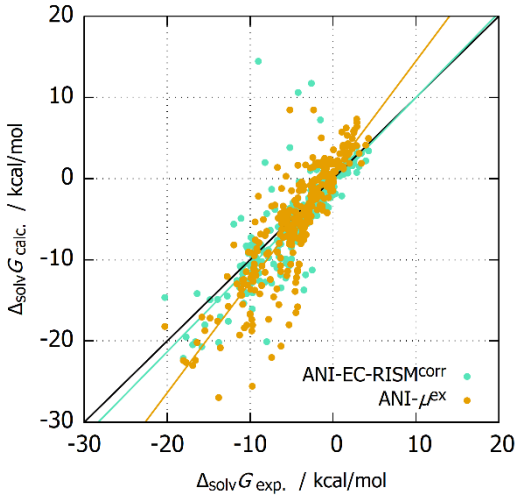

Figure 36: Calculated vs. experimental solvation free energies of the neutral molecules of the MNSOL database[196] using the ANI-EC-RISM^{corr}//B3LYP-6-311+G**/PCM and ANI-μ^{ex}//B3LYP-6-311+G**/PCM level of theory. Geometries were provided by M. Strobl.[275] A subset of 320 molecules was used for which no formally charged atoms were present. For ANI-μ^{ex} a PMV correction with a C_v of -0.07708 kcal/mol/Å^3 was performed using partial molar volumes calculated on the MP2/6-311+G**/EC-RISM level of theory also provided by M.Strobl.[275] Raw data is published in the electronic supplementary material, see chapter 6.1.

Table 37: Root mean squared error (RMSE), mean signed error (MSE) and mean absolute error (MAE) in kcal/mol compared to experimental solvation free energies. Regression parameters *m* and *b* are given for the linear regression together with the R^2 coefficient. A subset of 320 molecules was used for which no formally charged atoms were present. The same set of geometries optimized on the B3LYP/6-311+G**/PCM level of theory and provided by M. Strobl was used for both methods.[275] For ANI-μ^{ex} a PMV correction with a C_v of -0.07708 kcal/mol/Å^3 was performed using partial molar volumes calculated on the MP2/6-311+G**/EC-RISM level of theory also provided by M.Strobl.[275]

Method	RMSE	MSE	MAE	*m*	*b*	R^2
ANI-EC-RISM^{corr}	3.05	-0.64	1.87	1.044	-0.451	0.693
ANI-μ^{ex}	3.82	-0.70	2.72	1.362	0.855	0.746

Comparing ANI-EC-RISM^{corr} and ANI-μ^{ex}, ANI-EC-RISM^{corr} yields the better RMSE of 3.05 kcal/mol compared to 3.82 kcal/mol for ANI-μ^{ex}. In addition, the slope and intercept of the regression through both methods is much better for ANI-EC-RISM^{corr} where ANI-μ^{ex} shows a systematic underestimation of the most hydrophilic compounds. The fitted C_v of -0.07708 kcal/mol/Å^3 also drastically differs compared to what was used for ANI-EC-RISM^{corr} which was derived from EC-RISM calculations at the r^{2}SCAN-3c level of theory. This shows that ANI-μ^{ex} is not a reliable model for predicting solvation free energies. The reason for this might be due to the small range of μ^{ex} during the training from which no distinguishable features get extracted varying μ^{ex}. This then shows a worse transferability and yields a worse prediction for a test set such as the MNSOL. It is noteworthy that the intrinsic uncertainty of ANI-μ^{ex} as shown in Table 36 was significantly lower and the increased errors shown here are systematic. Based on this finding ANI-μ^{ex} is an unreliable model which needs further improvement.

In contrast however, ANI-EC-RISM^{corr} does not show any systematic errors. This is further validated by the fact that an additional PMV correction of the data in Figure 36 performed for ANI-EC-RISM^{corr} (using the PMV data of MP2 calculations provided by M.Strobl) yields an additional C_v of only 0.00155 kcal/mol/Å^3. Applying this additional correction only decreases the RMSE to 3.04 kcal/mol compared to the original 3.05 kcal/mol. This shows, that the approach of applying the PMV correction directly to the training data works. The finished MLP then does not need any additional fitting and can directly be used. The required vacuum energies to calculate the solvation free energy are cheap to compute compared to the solution phase increasing the prediction speed. They could also be predicted by another MLP, however the computational speedup by predicting the gas phase energies would be much lower compared to solution phase.

Regarding the prediction quality Figure 36 shows some large but unsystematic outliers of ANI-EC-RISM^{corr} vastly worsening the overall prediction metrics. To further investigate the predictive quality of ANI-EC-RISM^{corr} a cumulative RMSE was calculated based on the ordered set of data sorted by the absolute error compared to experiments. Figure 37 shows the RMSE for an increasing percentage of molecules included in the ordered set. As a reference MP2/EC-RISM data provided by M. Strobl is shown for comparison. Here the full dataset of 374 neutral molecules was used, however both methods get normalized to the ratio of molecules included.

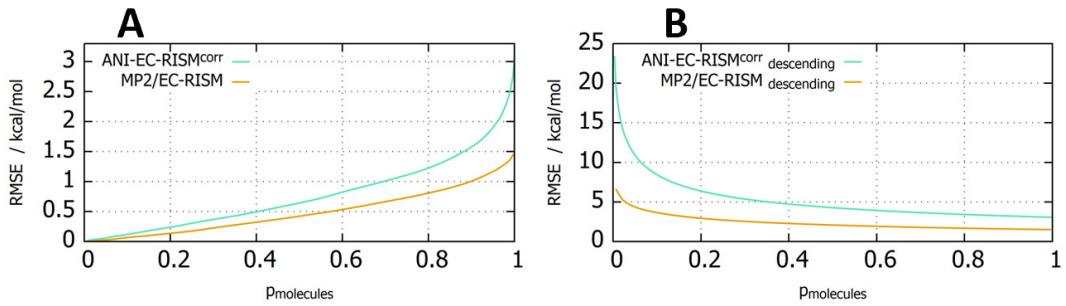

Figure 37: Running RMSE for an increasing number of datapoints of the calculated vs. experimental solvation free energies of the neutral molecules of the MNSOL database[196] using the ANI-EC-RISM^{corr}//B3LYP-6-311+G**/PCM and MP2-6-311+G**//B3LYP-6-311+G**/PCM ("MP2/EC-RISM") level of theory. A subset of 320 molecules was used for ANI-EC-RISM for which no formally charged atoms were present. The MP2/EC-RISM data is shown for all 374 neutral molecules of MNSOL database. MP2/EC-RISM data was provided by M. Strobl.[275] The RMSE is calculated for the part of molecules included "$p_{molecules}$". $p_{molecules}$ includes a growing number of molecules based on the absolute prediction error which in **A** was sorted in increasing order and in **B** in decreasing order. In this way the first point in **A** is the best prediction while in **B** this shows the worst prediction.

As expected, and by design, Figure 37, **A** shows that the RMSE increases with a larger fraction of molecules. However, for ANI-EC-RISM it drastically increases only for the last 10% of molecules (ca. 30 molecules) where the RMSE almost doubles from ca. 1.7 to 3.05 kcal/mol. In contrast using MP2/EC-RISM the increase over the last 10% is only from 1.0 to 1.5 kcal/mol. This is due to the fact that the worst ca. 10% of molecules have an overall RMSE of ca. 8 kcal/mol for ANI-EC-RISM^{corr} (Figure 37, **B**) and only ca. 4 kcal/mol for MP2/EC-RISM. Because the prediction is done by a neural network potential this hints at an incomplete training data set where specific molecules are underrepresented, leading to large prediction error of up to 23 kcal/mol while MP2/EC-RISM has a largest error of only 6.8 kcal/mol. The overall prediction quality of MP2/EC-RISM with an error of 1.5 kcal/mol can be reached for up to ca. 87% of molecules. Using the whole ANI-2x dataset for training might especially reduce the number and magnitude of these outliers, however this would also increase the computational cost for this model by a factor of ca. 20.

To further put the performance of the ANI-EC-RISM model into perspective a comparison to other methods is made in Table 38.

Table 38: Prediction error (RMSE) compared to experimentally measured solvation free energies of the MNSOL database. MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM data was provided by M. Strobl.[275]

Method	RMSE / kcal/mol
MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM	1.50
r^2SCAN-3c/EC-RISM// r^2SCAN-3c/CPCM[273]	1.50
r^2SCAN-3c/CPCM[273]	2.69
ANI-EC-RISM^{corr}// B3LYP/6-311+G**/PCM	3.05
ANI-μ^{ex}// B3LYP/6-311+G**/PCM	3.82

Here the ANI-EC-RISM^{corr} model is close to the CPCM model using the same level of theory. This shows that the more sophisticated solvation model, EC-RISM, that the MLP was trained to almost outweighs the intrinsic uncertainty and errors introduced by the MLP. Even compared to relatively cheap CPCM calculations the prediction by the ANI-EC-RISM model is close in accuracy to the r^2SCAN-3c/CPCM level of theory while still being orders of magnitude faster, which highlights the potential of the MLP, which could further be improved by a larger training dataset.

4.6.3 Benchmark of a solvated ANI-type Machine Learning Potential on the Tautobase dataset

As a second benchmark dataset the Tautobase[274] was used to evaluate the ANI-EC-RISM MLP. Here the tautomerization energy was calculated directly as the difference of G^{sol} which is predicted by ANI-EC-RISM and ANI-EC-RISM^{corr}. In addition, it can be calculated indirectly (see Figure 33) using ANI-μ^{ex} with gas phase data which here was provided on the MP2/6-311+G** level of theory by M. Strobl.[275] The impact of the PMV correction on tautomeric equilibria is expected to be small and therefore neglected for all calculations performed here. This also shows the potential of all methods to be used without a training to the MNSOL database as done in the previous chapter.

For the Tautobase an extensive outlier detection was performed by M. Strobl and used to create a filtered Tautobase removing 9 experimental outliers (IDs: 331, 735, 853, 1555, 1556, 1559, 1668, 1669 and 1670) which are disregarded throughout this chapter. For the high level MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM calculations this removal lowered the RMSE from 3.45 to 2.77 kcal/mol.[275]

To first evaluate the different models prediction metrics (RMSE, MSE and MAE) and regression parameters (m, b and R^2) are given in Table 39. A scatter plot of calculated and experimental reaction free energies for the tautomerization is given in Figure 38, **A**. Here the direct calculation with the r^2SCAN-3c/EC-RISM//B3LYP/6-311+G**/PCM level of theory is compared with the most similar MLP,

ANI-EC-RISM on the same geometries, provided by M. Strobl. All predictions within this chapter are made based on the lowest energy conformer found for each method.

To evaluate the potential use case of the ANI-EC-RISM MLP to optimize structures the respective minimum conformer was were used in high level MP2 calculations and compared to the standard workflow employing the B3LYP/6-311+G**/PCM level of theory for geometry optimizations. As a further comparison the ANI-2x[66] model was also employed for geometry optimizations. The geometry optimization was performed for multiple conformers for each tautomer but only the minimum conformer of each respective method was chosen for the MP2 calculation. As not all MP2 calculations converged for each set of geometries a subset of 367 molecules was analyzed for which all geometries lead to converging MP2 calculations. The resulting prediction metrics (RMSE, MSE and MAE) and regression parameters (m, b and R^2) are given in Table 39 too.

Table 39: Root mean squared error (RMSE), mean signed error (MSE) and mean absolute error (MAE) in kcal/mol for the tautomeric equilibria of 446 structures from the Tautobase using different theoretical methods. MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM results as well as B3LYP/6-311+G**/PCM geometries provided by M. Strobl.[275] For this method together with all new methods shown here a subset of 446 tautomer pairs was used after extensive outlier analysis by M. Strobl and applied also to methods from literature by M. Strobl. [275] For geometry optimizations using an ANI-type model the auto3d software was used.[77] The ANI-μ^{ex}// B3LYP/6-311+G**/PCM(indirect) method uses gas phase data on the MP2/6-311+G** level of theory provided by M. Strobl.[275] High level MP2/6-311+G**/EC-RISM calculations were performed on differently optimized geometries. For comparison a subset of 367 molecules was used for which all calculations converged. Raw data is published in the electronic supplementary material, see chapter 6.1.

Method	RMSE	MSE	MAE	*m*	*b*	R^2
B3LYP/6-31G(d)/SMD(indirect)[276]	2.62	0.39	2.02	1.126	0.401	0.853
MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM(direct)	2.77	1.20	2.16	1.129	1.218	0.864
Moltaut(direct)[277]	2.97	0.70	2.29	1.060	0.709	0.797
r^{2}SCAN-3c/EC-RISM// B3LYP/6-311+G**/PCM(direct)	3.57	-1.37	2.86	1.254	-1.329	0.833
ANI-EC-RISM^{corr}//ANI-EC-RISM^{corr}(direct)	4.18	-1.05	3.17	1.085	-1.052	0.683
ANI-EC-RISM//ANI-EC-RISM(direct)	4.30	-1.61	3.36	1.105	-1.592	0.696
ANI-EC-RISM^{corr}// B3LYP/6-311+G**/PCM(direct)	4.44	-0.95	3.31	1.085	-1.052	0.656
ANI-EC-RISM// B3LYP/6-311+G**/PCM(direct)	4.47	-1.62	3.51	1.187	-1.586	0.710
ANI-μ^{ex}// B3LYP/6-311+G**/PCM(indirect)	6.96	-1.38	5.38	1.584	-1.283	0.661
ANI-2x//ANI-2x(direct)	7.19	3.75	5.90	1.264	3.756	0.571
Subset of 367 molecules for which all calculations with the methods below converged						
MP2/6-311+G**/EC-RISM//B3LYP/6-311+G**/PCM(direct)	2.80	1.20	2.17	1.126	1.195	0.862
MP2/6-311+G**/EC-RISM//ANI-EC-RISM(direct)	3.30	1.35	2.59	1.170	1.340	0.829
MP2/6-311+G**/EC-RISM//ANI-2x(direct)	3.75	1.95	3.05	1.099	1.950	0.780

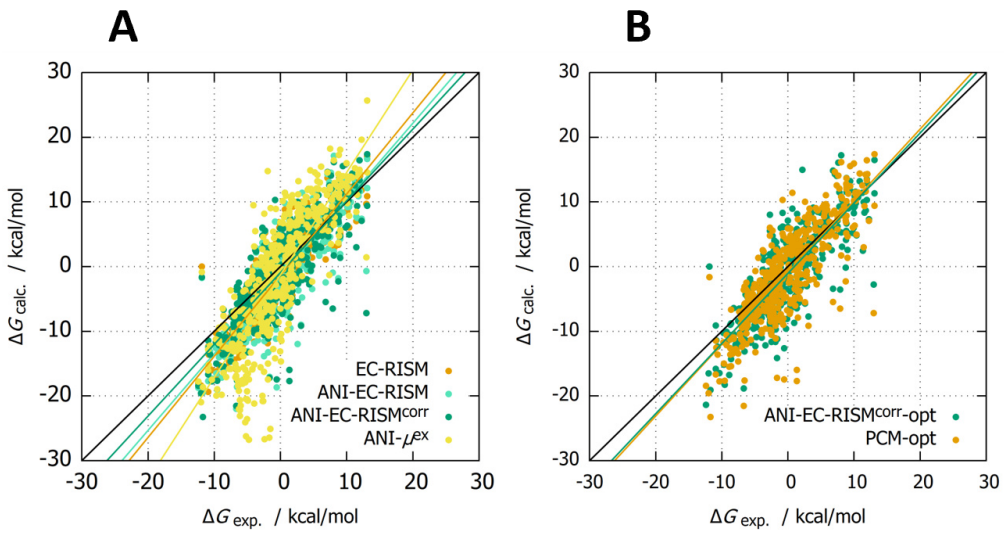

Figure 38: **A:** Calculated and experimental free energy differences in solution of tautomeric pairs for a subset of the Tautobase containing 446 structures (subset created by M. Strobl after extensive analysis of experimental outliers) using direct calculations with the r^{2}SCAN-3c/EC-RISM//B3LYP/6-311+G**/PCM level of theory ("EC-RISM") and the MLPs. For ANI-EC-RISM and ANI-EC-RISM^{corr} the direct calculation was performed, while for ANI-μ^{ex} the indirect calculation with MP2/6-311+G** gas phase data, provided by M.Strobl was performed.[275] All calculations in **A** used the same set of geometries. **B** shows a comparison between ANI-EC-RISM^{corr}- and PCM-geometries with the ANI-EC-RISM^{corr} model.

Figure 38 and Table 39 shows that ANI-EC-RISM and ANI-EC-RISM^{corr} are almost able to match the prediction quality of direct EC-RISM calculations at the same level of theory (r^{2}SCAN-3c). The additional error is only ca. 0.9 kcal/mol, which is lower than the uncertainty of the MLP itself (2.5 kcal/mol). This highlights the fact that these MLPs benefit from error compensation in the prediction of tautomeric equilibria. It is also noteworthy that the linear regression for both direct EC-RISM calculations and ANI-EC-RISM and ANI-EC-RISM^{corr} show very similar slopes and intercepts. From this it can be concluded that there is no remaining systematic error in the MLP besides the intrinsic error of the used level of theory. Errors due to the incomplete training data and the intrinsic uncertainty of the MLP are distributed evenly when compared to experiments. From this fact alone the training of these two MLPs can be deemed successful and the use in direct property predictions such as tautomeric equilibria can be recommended if the goal is a quick prediction for a large number of target molecules. Expectedly the absolute uncertainty of the predictions is slightly worse than the direct calculation, however the computational cost is orders of magnitude smaller.

In contrast the indirect approach of using gas phase calculations together with ANI-μ^{ex} yields high errors due to a systematic error with a too high a slope of more than 1.5. This is a similar behavior as found for the MNSOL (see Figure 36) and once again shows flaws for ANI-μ^{ex} most likely due to overfitting because of the small dynamic range of the training data. With that the use of ANI-μ^{ex} is not recommended for the calculation of tautomeric equilibria as it even requires the additional gas phase calculations.

The importance of the used geometries in the prediction of tautomeric equilibria with ANI-EC-RISM and ANI-EC-RISM^{corr} was also investigated. Here the prediction based on original geometries optimized on the B3LYP/6-311+G**/PCM level of theory (Table 39) were compared with the respective optimized structures. Overall, the RMSE decreases slightly from 4.47 kcal/mol for PCM-optimized structures to 4.30 kcal/mol for ANI-EC-RISM-optimized structures and from 4.44 kcal/mol to 4.18 kcal/mol for ANI-EC-RISM^{corr} (see Figure 38, **B**). This is mostly because several outliers get improved. As the RMSE against experimental data gets improved, this more consistent approach of using the minimum structure according to the MLPs potential energy surface yields better predictions. This is expected and should be the case as the PCM-optimized structures might not directly match the minimum of ANI-EC-RISM or ANI-EC-RISM^{corr}.

In comparison to methods from literature in Table 39 the best new model ("ANI-EC-RISM^{corr}//ANI-EC-RISM^{corr}") performs only slightly worse (ca. 1-1.5 kcal/mol) than other published methods. This is partly due to the fact that the training level of theory r^{2}SCAN-3c/EC-RISM is already inferior and the additional uncertainty due to the neural network decreases the prediction quality further. Compared to the ANI-2x model from which the training data was taken, ANI-EC-RISM is clearly superior in the prediction of tautomeric equilibria. This shows that the inclusion of solvent effects is a necessity for MLPs applied in solution. This is also true for tautomeric equilibria which seemed to drastically depend on solvent effects which must be included in the model.

Analyzing the performance of ANI-EC-RISM when used for geometry optimizations shows that it only introduces an additional error of 0.5 kcal/mol compared to the reference method B3LYP/6-311+G**/PCM (from 2.80 to 3.30 kcal/mol). As the single point calculations were all performed with high level MP2 calculations this error shows that ANI-EC-RISM geometries are slightly further away from the global optimum. This can be because a false conformer was chosen for the single point calculation, or the same conformer has worse geometric parameters such as bond lengths or angles. As the calculation of high-level data for all conformers optimized using these methods (more than 5000 for each method) was not feasible a definitive dissection of both potential error sources was not possible.

4.6.4 Summary and outlook

In summary the development of a solvated ANI-type MLP showed promising results. Even though only a small portion of the original ANI-2x dataset was used in the training of the model resulting errors on independent benchmark datasets (Tautobase and MNSOL) were close to the original EC-RISM reference on the same level of theory. Additional errors of less than 1 kcal/mol (from 3.57 to 4.44 kcal/mol) for tautomeric equilibria and less than 2 kcal/mol (from 1.50 to 3.05 kcal/mol) for solvation free energies are offset by prediction speeds three to four orders of magnitude faster. The inclusion of solvation effects with the EC-RISM solvation model is a necessity when applying the MLP

to phenomena in solution. This is highlighted by the fact that ANI-EC-RISM and ANI-EC-RISM^{corr} already outperform the full ANI-2x model on the prediction of the tautomeric equilibria and the geometry optimizations in solution with only 4% of the training data. Adding more training data would also likely further increase all prediction metrics compared to the ones shown here. In contrast the use of ANI-μ^{ex} is not recommended as it shows systematic errors.

While the computational demand is enormous to generate the training data for the ANI-EC-RISM potential use cases in the workflow for property prediction for a fully trained MLP are also widespread. From the results of the geometry optimization in which the ANI-EC-RISM geometries only introduced an additional error of 0.5 kcal/mol compared to the often used B3LYP/6-311+G**/PCM level of theory a fully trained ANI-EC-RISM potential might even be able to outperform the lower level methods used for geometry optimizations. In general, the here developed ANI-EC-RISM and ANI-EC-RISM^{corr} MLPs can be used for all applications that benefit from a fast prediction that does not need the highest accuracy. This could be a prefiltering of large sets of molecules with respect to a given property of interest, or a prefiltering of conformers or tautomers for a given molecule. As multiple tautomers or conformers might need to be evaluated for each molecule the reduction in computational cost would still be large.

High level single point calculations on the optimum determined by an ANI-EC-RISM MLP would still be necessary for high quality predictions as it is not feasible to fully remove the intrinsic uncertainty of the MLP and even the fully trained ANI-2x model had a validation RMSE of 1.8 kcal/mol[66] but ensuring that these calculations are performed on a closer to optimal structure would still improve the overall prediction quality. The geometry optimizations themselves would also be orders of magnitude faster with the possibility already to optimize the structure of small molecules within a few seconds using a single GPU.

Apart from the here developed ANI-EC-RISM potential of water at ambient conditions the success of the potentials also opens up the possibility of further potentials with all of the possible options available with EC-RISM. This includes the use of other organic solvents, such as Octanol or Cyclohexane which have previously been successfully treated with EC-RISM.[174][177]With improvements in hardware and therefore computational speedup it would also be possible to apply the ANI-EC-RISM potential to high-pressure conditions.

Something that was not investigated with ANI-EC-RISM is the application in MD simulations. As shown in earlier chapters the ANI-type MLPs can be applied to MD simulations and show improved structural parameters compared to classical FFMD. An ANI-EC-RISM potential would be able to sample the conformational space without the need for explicit solvent which takes up most of the computational demand of an MD simulation. In this approach ANI-EC-RISM could be comparable to implicit MD simulation methods such as the GB solvent model.[71][72] With the improved level of theory of EC-RISM compared to other implicit solvent models it is likely that an ANI-EC-RISM potential would outperform the GB solvent model.

To realize an ANI-EC-RISM potential suitable for MD simulations additional training data for atomic forces would be required as calculating forces during the simulation would be restrictive for potential use cases. Within EC-RISM however, calculating additional atomic forces would not require much additional calculation time, as it does not require the full EC-RISM cycle and can be done on the final solvated wave function from the EC-RISM calculation.

5. Overall summary and outlook

To summarize the work of this thesis a combined theoretical and experimental investigation was performed to successfully disentangle the quantitative structural and solvation response of NMR and EPR parameters.

It was shown that NMR experiments are highly accurate and provide ideal benchmark data to improve theoretical methods even under extreme conditions such as high hydrostatic pressure. Experimental investigations showed a previously unknown pressure-dependence of the ^{13}C reference compound DSS which must be included in the same way as it is done within theoretical methods to yield reliable insights from theoretical methods.

Disentangling structural and solvation response was first done by performing ensemble NMR calculations based on AIMD snapshots. These reached the accuracy limit with 4 Å of explicit solvent around the nucleus of interest. From this, sweet spots of accuracy and computational cost were identified with a local explicit solvation approach together with EC-RISM background solvation yielding a still high accuracy at much lower computational cost. This approach is especially interesting for all investigations that rely on NMR data for a single nucleus.

Replacing the AIMD simulations by cheaper FFMD simulations turned out to be more difficult depending on the molecule in question. Here the use of tailormade force fields, such as here for TMAO, still yields accurate results and is recommended whenever these force fields are available. However, using general force fields can introduce larger errors as shown here for the amide protons, aromatic carbons and ammonia.

The use of MLPs can improve on these as shown here for the case of NMA. However, more work is necessary to be able to use these MLPs for molecules such as DSS that could not be simulated at this point. Even the approach of starting from an FFMD ensemble and optimizing this by very short MLMD simulations turned out to improve all NMR parameters and could be an interesting approach but needs further investigation for a broad range of molecules.

Moving over to the high-pressure regime newly developed EC-RISM geometry optimizations were used to investigate the solute structure change under high pressure. It turned out that high pressure only has a marginal impact on bond lengths even with the sign of the change depending on the specific bond, but this was still enough to significantly impact the pressure-dependent change of chemical shifts. This shows that inclusion of these small structural changes is a necessity to accurately predict chemical shift changes under pressure.

Even though EC-RISM geometry optimizations include pressure-dependent geometry changes they showed their prowess already at ambient conditions outperforming PCM-geometries and even explicit solvation methods based on FFMD simulations. These simulations however were the best performing method for the pressure-dependent chemical shift changes as the only method which produces a

positive slope for the linear regression for ^{1}H, ^{13}C and ^{15}N. The last nucleus being still in question due to an uncertain experimental reference which needs further investigation. Testing the AIMD or MLMD in terms of high-pressure simulations would also be an interesting study even only for a ^{13}C and ^{15}N reference compound which were identified as the most significant error sources from FFMD simulations. If these simulation methods are not available due to technical limitations, the EC-RISM geometry optimizations are still the recommended choice for NMR predictions as general force fields do not yield sufficiently accurate solute structures.

With accurate chemical shift predictions, it is then possible to even calculate thermodynamic properties such as the pressure-dependent conformer equilibrium from reliable spectroscopic data, which was here performed for two dipeptides Ac-Ala-NHMe and Ac-Gly-NHMe. Using EC-RISM geometry optimizations together with a fitting on an independent set of amide protons confirmed the small intrinsic pressure-dependence of the protein backbone, found independently from direct energetic calculations. The accuracy of these thermodynamic predictions however relies on the accuracy of the spectroscopic predictions, which is why a large part of this thesis was devoted to develop methods and best practices to accomplish these predictions.

Here not only NMR parameters but also EPR parameters were analyzed. In another project the origin of the pH-dependence of the EPR spin probe HMI was quantitatively determined. By once again varying the solvent method used during EPR calculations it was possible to determine that 80% of the experimental change in both the g^{iso} and A^{iso} can be attributed to a change in electronic structure between the protonated $HMIH^{+}$ and the neutral HMI. The remaining 20% are due to the change in solvation which manifests as a change in the number of hydrogen bonds to the EPR-active nitroxide of HMI. Further investigation of the theoretical methods showed that EC-RISM background solvation in addition to local explicit solvation is the preferred method for quantitatively accurate EPR calculations outperforming QM/MM background solvation due to the larger solvent box modelled within EC-RISM as long-range solvent interactions still impact the resulting EPR parameters.

Lastly, the EC-RISM solvation model was used to create a general-purpose MLP trained on calculations in water. This model is then primarily used to directly access thermodynamic properties but it could also be used in implicit solvation MD simulations for ensemble-NMR calculations improving on FFMD simulations.

In comparison to direct EC-RISM calculations the best developed models (ANI-EC-RISM and ANI-EC-RISM^{corr}) introduced an additional RMSE of less than 1 kcal/mol in the predictions of tautomeric equilibria with respect to experimental measurements. This error is slightly larger for the prediction of solvation free energies with an additional error of ca. 1.5 kcal/mol compared to direct EC-RISM calculations. However, using the machine learning hundreds of energy calculations can be performed within a few seconds compared to hundreds of days of CPU-time for the direct calculation making this approach especially interesting for the prefiltering of large datasets with respect to a target property.

The MLP was also tested for geometry optimizations and yielded an additional error of ca. 0.5 kcal/mol compared to the standard workflow using geometry optimizations in implicit solvent. In both the

direct prediction of energetic quantities and the use for geometry optimizations the EC-RISM-based MLP outperformed established MLPs from literature which were trained on gas phase data and have more than 20 times more training data.

This highlights the importance of using high-quality solution phase data to train MLPs that are aimed to solve research questions in fields such as drug discovery or material designs which happen in solution. With the developed MLP it is also possible to run implicit solvent MD simulations even on complex systems such as proteins. Here the use of the ANI-type architecture allows for fast calculations making large scale MD simulations on a variety of targets viable. The use of the implicit solvent model here yields a systematic speedup showing that this approach improves in a variety of methods compared to existing MLPs relying on an explicit solvation approach.

The success of the EC-RISM-based MLP also opens the possibility to develop similar potentials in other solvents or solvent mixtures or high-pressure conditions where EC-RISM was already successfully used. The possibilities are only limited by the quality and quantity of the training data that can be generated.

6. Appendix

6.1 Description of electronic research data

All raw data of this thesis is provided electronically in machine readable format. The details of what is deposited are provided here:

Two research data publications have been made already which include:

1. https://doi.org/10.17877/RESOLV-2025-M7UFWH5H

Snapshots from AIMD simulations (performed by B. Sharma) of 2,2,3,4,5,5-hexamethylimidazolidin-1-oxyl in its neutral (HMI) and protonated form (HMIH+) in water with corresponding calculated EPR parameters (g-tensor and ^{14}N Hyperfine coupling constant A) using the following solvation methods:

* Vertically desolvated (VD) structures containing just HMI.

* HMI with an explicit second solvation shell (SSS) taken from AIMD.

* HMI with the explicit second solvation shell and an additional EC-RISM background (EC-RISM).

* HMI with the explicit second solvation shell and an additional QM/MM backround (QM/MM).

For EC-RISM and QM/MM the point charges necessary to reproduce the EPR calculation are included. Optimized structures and corresponding energies for diastereomers of HMIH+ that validate the simulated diastereomer as the main diastereomer of HMIH+ are included. The dataset contains example inputs to reproduce all energy and EPR calculations shown in chapter 4.5.

2. https://doi.org/10.17877/RESOLV-2024-lrgkievk:

* Snapshots from AIMD simulations (performed by B. Sharma) of DSS, NMA and TMAO with corresponding shielding constants shown in chapter 4.2.1 and 4.2.2.

* Optimized structures of DSS, NMA and TMAO with corresponding shielding constants shown in chapter 4.2.1 and 4.2.2.

* Example inputs to reproduce the calculated shielding constants including all further MD based NMR calculations including MLMD based NMR calculations shown in chapter 4.2.4 and 4.2.5 and pressure-dependent MD based NMR calculations shown in chapter 4.4.1.

6. Appendix

Further research data is newly published at:

3. https://doi.org/10.17877/RESOLV-2025-MJSL8MTK

This includes:

* Experimental pressure-dependent NMR measurements with the experimental FID and acquisition parameters used.

* Structures (including point charges) from MLMD and corresponding shielding constants to reproduce data from chapter 4.2.4 and 4.2.5.

* Pressure dependent structures and corresponding shielding constants to reproduce data from chapter 4.1.2 and 4.4.1.

* Pressure dependent structures and corresponding magnetic susceptibilities to reproduce data from chapter 4.1.3.

* ANI-EC-RISM-type datasets developed in chapter 4.6.

* Structures and corresponding energies to reproduce data from chapter 4.6.

6.2 Raw data for the effect of ensemble calculations on NMR parameters

Table 40: 1H and 13C shielding constants / ppm of DSS with "local" explicit solvent molecules and additional PCM or EC-RISM background for the AIMD ensemble. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent				
r_{water}	N_{water}	Solvent background	σ_H	σ_C
-	0	-	31.74 ± 0.01	185.40 ± 0.13
2.5 Å	5.4	-	31.60 ± 0.02	183.68 ± 0.14
3.0 Å	14.5	-	31.44 ± 0.02	182.49 ± 0.18
4.0 Å	27.9	-	31.28 ± 0.02	182.22 ± 0.13
4.5 Å	35.2		31.28 ± 0.03	182.23 ± 0.14
5.0 Å	44.9		31.28 ± 0.02	182.20 ± 0.13
5.5 Å	56.9		31.26 ± 0.01	182.22 ± 0.13
-	0	PCM	31.64 ± 0.01	186.10 ± 0.12
2.5 Å	5.4	PCM	31.50 ± 0.01	184.37 ± 0.13
3.0 Å	14.5	PCM	31.34 ± 0.01	183.18 ± 0.13
4.0 Å	27.9	PCM	31.21 ± 0.01	182.70 ± 0.13
4.5 Å	35.2	PCM	31.20 ± 0.01	182.65 ± 0.13
5.0 Å	44.9	PCM	31.20 ± 0.01	182.63 ± 0.13
5.5 Å	56.9	PCM	31.20 ± 0.01	182.60 ± 0.13
-	0	EC-RISM	31.62 ± 0.01	186.53 ± 0.14
2.5 Å	5.4	EC-RISM	31.48 ± 0.01	184.71 ± 0.14
3.0 Å	14.5	EC-RISM	31.30 ± 0.01	183.17 ± 0.14
4.0 Å	27.9	EC-RISM	31.19 ± 0.01	182.60 ± 0.13
4.5 Å	35.2	EC-RISM	31.18 ± 0.01	182.49 ± 0.13
5.0 Å	44.9	EC-RISM	31.18 ± 0.01	182.46 ± 0.13
5.5 Å	56.9	EC-RISM	31.18 ± 0.01	182.45 ± 0.13

Table 41: 1H and 13C shielding constants / ppm of DSS with "full" explicit solvent molecules and additional PCM or EC-RISM background for the AIMD ensembles. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent				
r_{water}	N_{water}	Solvent background	σ_H	σ_C
-	0	-	31.74 ± 0.01	185.40 ± 0.13
2.5 Å	13.3	-	31.51 ± 0.01	183.62 ± 0.10
3.0 Å	25.4	-	31.34 ± 0.01	182.57 ± 0.10
4.0 Å	43.3	-	31.22 ± 0.01	182.24 ± 0.09
-	0	PCM	31.64 ± 0.01	186.10 ± 0.12
2.5 Å	13.3	PCM	31.46 ± 0.01	184.26 ± 0.13
3.0 Å	25.4	PCM	31.30 ± 0.01	183.10 ± 0.13
4.0 Å	43.3	PCM	31.19 ± 0.01	182.67 ± 0.13
-	0	EC-RISM	31.62 ± 0.01	186.53 ± 0.12
2.5 Å	13.3	EC-RISM	31.45 ± 0.01	184.57 ± 0.13
3.0 Å	25.4	EC-RISM	31.27 ± 0.01	183.15 ± 0.13
4.0 Å	43.3	EC-RISM	31.16 ± 0.01	182.54 ± 0.13
-	0	QM/MM	31.66 ± 0.01	185.99 ± 0.13
2.5 Å	13.3	QM/MM	31.47 ± 0.01	184.08 ± 0.13
3.0 Å	25.4	QM/MM	31.29 ± 0.01	182.77 ± 0.13
4.0 Å	43.3	QM/MM	31.18 ± 0.01	182.33 ± 0.13

Table 42: Amide 1H shielding constants / ppm of NMA in the trans conformation with "local" explicit solvent molecules and additional PCM or EC-RISM background for the AIMD ensembles. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent			
r_{water}	N_{water}	Solvent background	σ_H
-	0	-	26.64 ± 0.04
2.5 Å	1.1	-	25.01 ± 0.05
3.0 Å	2.1	-	25.02 ± 0.06
4.0 Å	7.0	-	24.82 ± 0.07
-	0	PCM	25.84 ± 0.04
2.5 Å	1.1	PCM	24.33 ± 0.07
3.0 Å	2.1	PCM	24.32 ± 0.07
4.0 Å	7.0	PCM	24.16 ± 0.07
-	0	EC-RISM	24.65 ± 0.04
2.5 Å	1.1	EC-RISM	23.58 ± 0.07
3.0 Å	2.1	EC-RISM	23.62 ± 0.07
4.0 Å	7.0	EC-RISM	23.61 ± 0.07

Table 43: Amide 1H shielding constants / ppm of NMA in the trans conformation with "full" explicit solvent molecules and additional PCM, EC-RISM or QM/MM background for the AIMD ensembles. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent			
r_{water}	N_{water}	Solvent background	σ_H
-	0	-	26.64 ± 0.04
2.5 Å	6.7	-	24.53 ± 0.07
3.0 Å	14.8	-	24.34 ± 0.07
4.0 Å	28.3	-	24.05 ± 0.07
-	0	PCM	25.84 ± 0.04
2.5 Å	6.7	PCM	24.01 ± 0.07
3.0 Å	14.8	PCM	23.90 ± 0.07
4.0 Å	28.3	PCM	23.73 ± 0.07
-	0	EC-RISM	24.65 ± 0.04
2.5 Å	6.7	EC-RISM	23.78 ± 0.07
3.0 Å	14.8	EC-RISM	23.78 ± 0.07
4.0 Å	28.3	EC-RISM	23.73 ± 0.07
-	0	QM/MM	25.32 ± 0.05
2.5 Å	6.7	QM/MM	24.19 ± 0.07
3.0 Å	14.8	QM/MM	24.07 ± 0.07
4.0 Å	28.3	QM/MM	23.90 ± 0.07

Table 44: Amide 1H shielding constants / ppm of NMA in the cis conformation with "local" explicit solvent molecules and additional PCM or EC-RISM background for the AIMD ensembles. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent			
r_{water}	N_{water}	Solvent background	σ_H
-	0	-	26.77 ± 0.03
2.5 Å	1.3	-	25.23 ± 0.05
3.0 Å	2.9	-	25.11 ± 0.05
4.0 Å	8.2	-	24.89 ± 0.05
-	0	PCM	26.27 ± 0.04
2.5 Å	1.3	PCM	24.90 ± 0.07
3.0 Å	2.9	PCM	24.81 ± 0.07
4.0 Å	8.2	PCM	24.62 ± 0.07
-	0	EC-RISM	25.19 ± 0.04
2.5 Å	1.3	EC-RISM	24.31 ± 0.07
3.0 Å	2.9	EC-RISM	24.39 ± 0.07
4.0 Å	8.2	EC-RISM	24.41 ± 0.07

Table 45: Amide 1H shielding constants / ppm of NMA in the cis conformation with "full" explicit solvent molecules and additional PCM, EC-RISM or QM/MM background for the AIMD ensembles. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent			
r_{water}	N_{water}	Solvent background	σ_H
-	0	-	26.77 ± 0.04
2.5 Å	6.5	-	24.92 ± 0.08
3.0 Å	14.7	-	24.86 ± 0.07
4.0 Å	27.9	-	24.69 ± 0.08
-	0	PCM	26.27 ± 0.02
2.5 Å	6.5	PCM	24.68 ± 0.07
3.0 Å	14.7	PCM	24.62 ± 0.07
4.0 Å	27.9	PCM	24.49 ± 0.07
-	0	EC-RISM	25.19 ± 0.04
2.5 Å	6.5	EC-RISM	24.46 ± 0.07
3.0 Å	14.7	EC-RISM	24.48 ± 0.07
4.0 Å	27.9	EC-RISM	24.45 ± 0.08
-	0	QM/MM	25.72 ± 0.05
2.5 Å	6.5	QM/MM	24.72 ± 0.08
3.0 Å	14.7	QM/MM	24.64 ± 0.07
4.0 Å	27.9	QM/MM	24.51 ± 0.08

Table 46: 1H and 13C shielding constants / ppm of TMAO with "local" explicit solvent molecules and additional PCM or EC-RISM background. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent				
r_{water}	N_{water}	Solvent background	σ_H	σ_C
-	0	-	28.46 ± 0.01	122.15 ± 0.16
2.5 Å	5.7	-	28.24 ± 0.02	122.78 ± 0.19
3.0 Å	14.9	-	28.02 ± 0.02	122.28 ± 0.17
4.0 Å	25.0	-	27.90 ± 0.01	121.80 ± 0.16
-	0	PCM	28.42 ± 0.01	121.23 ± 0.17
2.5 Å	5.7	PCM	28.16 ± 0.02	122.12 ± 0.17
3.0 Å	14.9	PCM	27.95 ± 0.02	122.05 ± 0.22
4.0 Å	25.0	PCM	27.83 ± 0.01	121.61 ± 0.16
-	0	EC-RISM	28.32 ± 0.01	121.66 ± 0.17
2.5 Å	5.7	EC-RISM	28.10 ± 0.01	122.12 ± 0.17
3.0 Å	14.9	EC-RISM	27.93 ± 0.01	121.75 ± 0.17
4.0 Å	25.0	EC-RISM	27.82 ± 0.01	121.53 ± 0.16

Table 47: 1H and 13C shielding constants / ppm of TMAO with "full" explicit solvent molecules and additional PCM or EC-RISM background. The average number of water molecules corresponding to the applied distance criterion is given as well. Errors were calculated as the standard error over all snapshots. The shieldings are depicted in Figure 14 and were used to calculate the chemical shifts shown in Figure 15.

Explicit solvent				
r_{water}	N_{water}	Solvent background	σ_H	σ_C
-	-	-	28.45 ± 0.01	122.09 ± 0.16
2.5 Å	7.1 H_2O	-	28.17 ± 0.02	123.07 ± 0.17
3.0 Å	14.9 H_2O	-	28.03 ± 0.01	122.38 ± 0.17
4.0 Å	28.3 H_2O	-	27.91 ± 0.01	121.81 ± 0.16
-	-	PCM	28.40 ± 0.01	121.14 ± 0.17
2.5 Å	7.1 H_2O	PCM	28.09 ± 0.01	122.50 ± 0.17
3.0 Å	14.9 H_2O	PCM	27.96 ± 0.01	121.98 ± 0.16
4.0 Å	28.3 H_2O	PCM	27.84 ± 0.01	121.70 ± 0.16
-	-	EC-RISM	28.31 ± 0.01	121.56 ± 0.17
2.5 Å	7.1 H_2O	EC-RISM	28.05 ± 0.01	122.46 ± 0.17
3.0 Å	14.9 H_2O	EC-RISM	27.94 ± 0.01	121.87 ± 0.17
4.0 Å	28.3 H_2O	EC-RISM	27.83 ± 0.01	121.64 ± 0.16
-	-	QM/MM	28.36 ± 0.01	121.94 ± 0.17
2.5 Å	7.1 H_2O	QM/MM	28.10 ± 0.02	122.62 ± 0.17
3.0 Å	14.9 H_2O	QM/MM	27.97 ± 0.01	122.11 ± 0.16
4.0 Å	28.3 H_2O	QM/MM	27.83 ± 0.02	121.50 ± 0.17

6.3 Raw data for pressure-dependent NMR spectroscopy

Table 48: Pressure-dependent experimental chemical shifts of urea. ^{1}H and ^{13}C chemical shifts were directly referenced to internal DSS. The ^{15}N chemical shifts were referenced to the pressure-dependent ^{1}H signal of DSS using a pressure-independent Ξ of 0.25144953.[206] The experimental chemical shifts at 10 bar were used in comparison with ambient pressure calculations. Chemical shifts and chemical shift differences from 3000 bar to 1 bar are depicted in Figure 25.

p / bar	$\delta_{\mathbf{N}\mathrm{H}}$ / ppm	δ_{CO} / ppm	$\delta_{\mathbf{N}\mathrm{H}}$ / ppm
10	5.7699	165.50	76.84
500	5.7876	165.34	77.20
1000	5.8010	165.18	77.55
1500	5.8166	165.04	77.85
2000	5.8316	164.90	78.15
2500	5.8461	164.77	78.43
3000	5.8597	164.84	78.69

Table 49: Pressure-dependent experimental chemical shifts of acetamide. ^{1}H and ^{13}C chemical shifts were directly referenced to internal DSS. The ^{15}N chemical shifts were referenced to the pressure-dependent ^{1}H signal of DSS using a pressure-independent Ξ of 0.25144953.[206] The experimental chemical shifts at 10 bar were used in comparison with ambient pressure calculations. Chemical shifts and chemical shift differences from 3000 bar to 1 bar are depicted in Figure 25.

p / bar	$\delta_{\text{H-methyl}}$ / ppm	$\delta_{\mathbf{N}\mathrm{H}\text{-trans}}$ / ppm	$\delta_{\mathbf{N}\mathrm{H}\text{-cis}}$ / ppm	$\delta_{\text{C-methyl}}$ / ppm	$\delta_{C\text{-carbonyl}}$ / ppm	$\delta_{\mathbf{N}\mathrm{H}}$ / ppm
10	1.9930	7.5434	6.7824	24.0543	180.13	112.15
500	1.9943	7.5506	6.8055	24.0479	180.00	112.53
1000	1.9952	7.5584	6.8266	24.0408	179.88	112.90
1500	1.9962	7.5684	6.8452	24.0330	179.76	113.23
2000	1.9976	7.5770	6.8633	24.0259	179.65	113.55
2500	1.9987	7.5873	6.8799	24.0169	179.55	113.90
3000	1.9991	7.5969	6.8959	24.0075	179.45	114.22

Table 50: Pressure-dependent experimental chemical shifts of NMA. ^{1}H and ^{13}C chemical shifts were directly referenced to internal DSS. The ^{15}N chemical shifts were referenced to the pressure-dependent ^{1}H signal of DSS using a pressure-independent Ξ of 0.25144953.[206] The experimental chemical shifts at 10 bar were used in comparison with ambient pressure calculations. Chemical shifts and chemical shift differences from 3000 bar to 1 bar are depicted in Figure 25.

p / bar	$\delta_{\text{H-methyl-1}}$ / ppm	$\delta_{\text{NH-trans}}$ / ppm	$\delta_{\text{H-methyl-2}}$ / ppm	$\delta_{\text{NH-cis}}$ / ppm	$\delta_{\text{C-methyl-1}}$ / ppm	$\delta_{\text{C-carbonyl}}$ / ppm	$\delta_{\text{C-methyl-2}}$ / ppm	δ_{NH} / ppm
10	1.9672	7.8169	2.7024	7.0860	24.39	177.35	28.70	112.63
500	1.9697	7.8205	2.7006	7.1034	24.37	177.22	28.68	112.86
1000	1.9716	7.8248	2.6995	7.1193	24.36	177.09	28.65	113.08
1500	1.9732	7.8294	2.6970	7.1333	24.34	176.97	28.63	113.29
2000	1.9749	7.8346	2.6949	7.1497	24.33	176.86	28.61	113.48
2500	1.9762	7.8423	2.6929	7.1585	24.31	176.76	28.58	113.66
3000	1.9774	7.8463	2.6915	7.1696	24.30	176.66	28.56	113.83

Table 51: Pressure-dependent experimental ^{1}H chemical shifts of benzamide. Chemical shifts were directly referenced to internal DSS. The experimental chemical shifts at 10 bar were used in comparison with ambient pressure calculations. Chemical shifts and chemical shift differences from 3000 bar to 1 bar are depicted in Figure 25.

p / bar	$\delta_{\text{H-para}}$ / ppm	$\delta_{\text{H-meta}}$ / ppm	$\delta_{\text{H-ortho}}$ / ppm	$\delta_{\text{NH-trans}}$ / ppm	$\delta_{\text{NH-cis}}$ / ppm
10	7.6388	7.5337	7.8132	8.0499	7.2270
500	7.6407	7.5353	7.8129	8.0610	7.2489
1000	7.6426	7.5368	7.8126	8.0716	7.2691
1500	7.6447	7.5379	7.8124	8.0818	7.2870
2000	7.6469	7.5397	7.8123	8.1046	7.3059
2500	7.6491	7.5412	7.8122	8.1131	7.3211
3000	7.6518	7.5430	7.8122	8.1133	7.3374

Table 52: Pressure-dependent experimental ^{13}C and ^{15}N chemical shifts of benzamide. ^{13}C chemical shifts were directly referenced to internal DSS. The ^{15}N chemical shifts were referenced to the pressure-dependent ^{1}H signal of DSS using a pressure-independent Ξ of 0.25144953.[206] The experimental chemical shifts at 10 bar were used in comparison with ambient pressure calculations. Chemical shifts and chemical shift differences from 3000 bar to 1 bar are depicted in Figure 25.

p / bar	$\delta_{\text{C-para}}$ / ppm	$\delta_{\text{C-meta}}$ / ppm	$\delta_{\text{C-ortho}}$ / ppm	$\delta_{\text{C-ipso}}$ / ppm	$\delta_{\text{C-carbonyl}}$ / ppm	δ_{NH2} / ppm
10	135.24	131.45	130.09	135.34	175.90	106.77
500	135.14	131.34	130.00	135.13	175.73	107.12
1000	135.05	131.24	129.91	134.94	175.56	107.43
1500	134.96	131.13	129.88	134.75	175.39	107.73
2000	134.88	131.04	129.74	134.58	175.24	108.01
2500	134.80	130.95	129.66	134.41	175.08	108.27
3000	134.72	130.87	129.59	134.25	174.94	108.51

7. References

[1] O. A. El Seoud, *Pure Appl. Chem.* **2009**, *81*, 697.

[2] E. Paulus, *Front. Mar. Sci.* **2021**, *8*, 667048.

[3] J.-M. Knop, S. Mukherjee, M. W. Jaworek, S. Kriegler, M. Manisegaran, Z. Fetahaj, L. Ostermeier, R. Oliva, S. Gault, C. S. Cockell et al., *Chem. Rev.* **2023**, *123*, 73.

[4] E. Napiórkowska, K. Milcarz, Ł. Szeleszczuk, *Int. J. Mol. Sci.* **2023**, *24*.

[5] J. Eeckhoudt, M. Alonso, P. Geerlings, F. de Proft, *J. Chem. Theory Comput.* **2024**, *20*, 7430.

[6] H. R. Kalbitzer, I. C. Rosnizeck, C. E. Munte, S. P. Narayanan, V. Kropf, M. Spoerner, *Angew. Chem. Int. Ed.* **2013**, *52*, 14242.

[7] R. Frach, P. Kibies, S. Böttcher, T. Pongratz, S. Strohfeldt, S. Kurrmann, J. Koehler, M. Hofmann, W. Kremer, H. R. Kalbitzer et al., *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 8757.

[8] T. Pongratz, P. Kibies, L. Eberlein, N. Tielker, C. Hölzl, S. Imoto, M. Beck Erlach, S. Kurrmann, P. H. Schummel, M. Hofmann et al., *Biophys. Chem.* **2020**, *257*, 106258.

[9] B. Sharma, V. A. Tran, T. Pongratz, L. Galazzo, I. Zhurko, E. Bordignon, S. M. Kast, F. Neese, D. Marx, *J. Chem. Theory Comput.* **2021**, *17*, 6366.

[10] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.

[11] H. Günther, *NMR spectroscopy*. *Basic principles, concepts and applications in chemistry,* 3. Aufl., Wiley-VCH, Weinheim, **2013**.

[12] A. M. Gronenborn, T. Polenova, *Chem. Rev.* **2022**, *122*, 9265.

[13] Y. Nishiyama, G. Hou, V. Agarwal, Y. Su, A. Ramamoorthy, *Chem. Rev.* **2023**, *123*, 918.

[14] G. Bifulco, P. Dambruoso, L. Gomez-Paloma, R. Riccio, *Chem. Rev.* **2007**, *107*, 3744.

[15] P. H. Willoughby, M. J. Jansma, T. R. Hoye, *Nat. Protoc.* **2014**, *9*, 643.

[16] I. Cortés, C. Cuadrado, A. Hernández Daranas, A. M. Sarotti, *Front. Nat. Prod.* **2023**, *2*, 1122426.

[17] H. R. Kalbitzer in *Subcellular biochemistry*, *Vol. 72*, Springer, Dordrecht, **2015**, S. 179–197.

[18] K. Akasaka, H. Yamada, *Methods Enzymol.* **2001**, *338*, 134.

[19] K. Akasaka, *Subcell Biochem.* **2015**, *72*, 707.

[20] R. Kitahara, S. Sareth, H. Yamada, E. Ohmae, K. Gekko, K. Akasaka, *Biochemistry* **2000**, *39*, 12789.

[21] D. D. Boehr, D. McElheny, H. J. Dyson, P. E. Wright, *Science* **2006**, *313*, 1638.

[22] B. Fuglestad, M. A. Stetz, Z. Belnavis, A. J. Wand, *Biophys. Chem.* **2017**, *231*, 39.

[23] R. Kitahara, K. Akasaka, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3167.

[24] P. A. Jennings, P. E. Wright, *Science* **1993**, *262*, 892.

[25] K. Pervushin, R. Riek, G. Wider, K. Wüthrich, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12366.

[26] R. Kitahara, K. Hata, H. Li, M. P. Williamson, K. Akasaka, *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *71*, 35.

[27] H. Li, H. Yamada, K. Akasaka, *Biochemistry* **1998**, *37*, 1167.

[28] E. Benassi, *J. Comput. Chem.* **2017**, *38*, 87.

[29] I. G. Shenderovich, *Chem. Methods* **2021**, *1*, 61.

[30] R. D. Cohen, J. S. Wood, Y.-H. Lam, A. V. Buevich, E. C. Sherer, M. Reibarkh, R. T. Williamson, G. E. Martin, *Molecules* **2023**, *28*.

[31] P. Cimino, L. Gomez-Paloma, D. Duca, R. Riccio, G. Bifulco, *Magn. Reson. Chem.* **2004**, *42 Spec no*, S26-33.

[32] C. Han, D. Zhang, S. Xia, Y. Zhang, *J. Chem. Theory Comput.* **2024**, *20*, 5250.

[33] T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B* **2008**, *112*, 4337.

[34] S. Suzen, H. Gurer-Orhan, L. Saso, *Molecules* **2017**, *22*.

[35] W. R. Hagen, *Dalton Trans.* **2006**, 4415.

[36] L. Hofmann, S. Ruthstein, *J. Phys. Chem. B* **2022**, *126*, 7486.

[37] H. Beinert, R. H. Sands, *Biochem. Biophys. Res. Commun.* **1960**, *3*, 41.

[38] W. L. Hubbell, H. S. Mchaourab, C. Altenbach, M. A. Lietzow, *Structure* **1996**, *4*, 779.

[39] J. P. Klare, *Biol. Chem.* **2013**, *394*, 1281.

[40] G. Z. Sowa, P. Z. Qin, *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *82*, 147.

[41] W. K. Subczynski, M. Raguz, J. Widomska, *Methods Mol. Biol.* **2010**, *606*, 247.

[42] D. Goldfarb, S. Stoll, *EPR spectroscopy*. *Fundamentals and methods*, Wiley, Chichester, West Sussex, **2018**.

[43] L. J. Berliner, J. Grunwald, H. O. Hankovszky, K. Hideg, *Anal. Biochem.* **1982**, *119*, 450.

[44] M. A. Voinov, A. I. Smirnov in *Specialist Periodical Report*, *Vol. 22*, Royal Society of Chemistry, Cambridge, **2010**, S. 71–106.

[45] M. A. Voinov, A. I. Smirnov, *Methods Enzymol.* **2015**, *564*, 191.

[46] R. Owenius, M. Engström, M. Lindgren, M. Huber, *J. Phys. Chem. A* **2001**, *105*, 10967.

[47] V. V. Khramtsov, L. M. Weiner, L. A. Grigoriev, L. B. Volodarsky, *Chem. Phys. Lett.* **1982**, *92*, 214.

[48] K. Margita, M. A. Voinov, A. I. Smirnov, *Cell Biochem. Biophys.* **2017**, *75*, 185.

[49] V. Barone, M. Fusè, S. M. V. Pinto, N. Tasinato, *Molecules* **2021**, *26*.

[50] O. I. Gromov, S. V. Kuzin, E. N. Golubeva, *J. Mol. Model.* **2019**, *25*, 93.

[51] Y. Zhao, D. G. Truhlar, *Org. Lett.* **2006**, *8*, 5753.

[52] A. A. Auer, A. van Tran, B. Sharma, G. L. Stoychev, D. Marx, F. Neese, *Mol. Phys.* **2020**, *118*, e1797916.

[53] Z. W. Windom, A. Perera, R. J. Bartlett, *J. Chem. Phys.* **2022**, *156*, 94107.

[54] G. Schreckenbach, T. Ziegler, *J. Phys. Chem. A* **1997**, *101*, 3388.

[55] Z. Rinkevicius, N. A. Murugan, J. Kongsted, K. Aidas, A. H. Steindal, H. Agren, *J. Phys. Chem. B* **2011**, *115*, 4350.

[56] J. Daru, H. Forbert, J. Behler, D. Marx, *Phys. Rev. Lett.* **2022**, *129*, 226001.

[57] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nat. Commun.* **2019**, *10*, 2903.

[58] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[59] I. Batatia, G. Csanyi, D. P. Kovacs, C. Ortner, G. Simm in *Advances in Neural Information Processing Systems 35*, Neural Information Processing Systems Foundation, Inc. (NeurIPS), San Diego, California, USA, **2022**, S. 11423–11436.

[60] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, *J. Chem. Phys.* **2020**, *153*, 34702.

[61] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Phys. Rev. X* **2018**, *8*.

[62] S. Ma, Z.-P. Liu, *Chem. Sci.* **2022**, *13*, 5055.

[63] A. M. Tokita, J. Behler, *J. Chem. Phys.* **2023**, *159*.

[64] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, A. R. Oganov, *Phys. Rev. B* **2019**, *99*.

[65] R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, R. Asahi, *J. Phys. Chem. Lett.* **2020**, *11*, 6946.

[66] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev, A. E. Roitberg, *J. Chem. Theory Comput.* **2020**, *16*, 4192.

[67] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole et al., *J. Am. Chem. Soc.* **2025**, *147*, 17598.

[68] L. Medrano Sandonas, D. van Rompaey, A. Fallani, M. Hilfiker, D. Hahn, L. Perez-Benito, J. Verhoeven, G. Tresadern, J. Kurt Wegner, H. Ceulemans et al., *Sci Data* **2024**, *11*, 742.

[69] J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.

[70] J. A. Semelak, I. Pickering, K. Huddleston, J. Olmos, J. S. Grassano, C. M. Clemente, S. I. Drusin, M. Marti, M. C. Gonzalez Lebrero, A. E. Roitberg et al., *J. Chem. Theory Comput.* **2025**.

[71] V. Tsui, D. A. Case, *Biopolymers* **2000**, *56*, 275.

[72] A. V. Onufriev, D. A. Case, *Annu. Rev. Biophys.* **2019**, *48*, 275.

[73] M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, O. Winther, *Sci Data* **2022**, *9*, 779.

[74] S. Zhang, R. Zubatyuk, Y. Yang, A. Roitberg, O. Isayev, *J. Chem. Theory Comput.* **2025**, *21*, 4365.

[75] D. L. Folmsbee, D. R. Koes, G. R. Hutchison, *J. Phys. Chem. A* **2021**, *125*, 1987.

[76] T. Ginex, J. Vázquez, C. Estarellas, F. J. Luque, *Curr. Opin. Struct. Biol.* **2024**, *87*, 102870.

[77] Z. Liu, T. Zubatiuk, A. Roitberg, O. Isayev, *J. Chem. Inf. Model.* **2022**, *62*, 5373.

[78] O. Tayfuroglu, I. N. Zengin, M. S. Koca, A. Kocak, *J. Chem. Inf. Model.* **2025**, *65*, 2818.

[79] P. A. M. Dirac, *Math. Proc. Camb. Phil. Soc.* **1939**, *35*, 416.

[80] M. Born, R. Oppenheimer, *Ann. Phys.* **1927**, *389*, 457.

[81] J. C. Slater, *Phys. Rev.* **1930**, *36*, 57.

[82] H. Taketa, S. Huzinaga, K. O-ohata, *J. Phys. Soc. Jpn.* **1966**, *21*, 2313.

[83] W. J. Hehre, R. F. Stewart, J. A. Pople, *J. Chem. Phys.* **1969**, *51*, 2657.

[84] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

[85] F. Jensen, *Introduction to Computational Chemistry,* 1. Aufl., John Wiley & Sons Incorporated, Newark, **2017**.

[86] W. Pauli, *Z. Physik* **1925**, *31*, 765.

[87] J. C. Slater, *Phys. Rev.* **1929**, *34*, 1293.

[88] D. R. Hartree, *Math. Proc. Camb. Phil. Soc.* **1928**, *24*, 111.

[89] V. Fock, *Z. Physik* **1930**, *61*, 126.

[90] P.-O. Löwdin, *Phys. Rev.* **1955**, *97*, 1509.

[91] R. Eisenschitz, F. London, *Z. Physik* **1930**, *60*, 491.

[92] C. Møller, M. S. Plesset, *Phys. Rev.* **1934**, *46*, 618.

[93] J. Řezáč, P. Hobza, *J. Chem. Theory Comput.* **2013**, *9*, 2151.

[94] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864-B871.

[95] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133-A1138.

[96] L. H. Thomas, *Math. Proc. Camb. Phil. Soc.* **1927**, *23*, 542.

[97] E. Fermi, *Rend. Accad. Naz. Lincei* **1927**, *6*, 32.

[98] P. A. M. Dirac, *Math. Proc. Camb. Phil. Soc.* **1930**, *26*, 376.

[99] R. G. Parr, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, Oxford, **2015**.

[100] M. Gell-Mann, K. A. Brueckner, *Phys. Rev.* **1957**, *106*, 364.

[101] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, Pederson, D. J. Singh, C. Fiolhais, *Phys. Rev. B Condens. Matter* **1992**, *46*, 6671.

[102] A. D. Becke, *Phys. Rev. A Gen. Phys.* **1988**, *38*, 3098.

[103] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, J. Sun, *J. Phys. Chem. Lett.* **2020**, *11*, 8208.

[104] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[105] S. Grimme, A. Hansen, J. G. Brandenburg, C. Bannwarth, *Chemical reviews* **2016**, *116*, 5105.

[106] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, D. C. Langreth, *Phys. Rev. B* **2010**, *82*.

[107] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.

[108] E. Caldeweyher, C. Bannwarth, S. Grimme, *J. Chem. Phys.* **2017**, *147*, 34112.

[109] O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger, D. Sebastiani, *Phys. Rev. Lett.* **2004**, *93*, 153004.

[110] W.-M. Hoe, A. J. Cohen, N. C. Handy, *Chem. Phys. Lett.* **2001**, *341*, 319.

[111] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B Condens. Matter* **1988**, *37*, 785.

[112] Y. Zhang, W. Yang, *Phys. Rev. Lett.* **1998**, *80*, 890.

[113] C. Adamo, V. Barone, *J. Chem. Phys.* **1999**, *110*, 6158.

[114] M. Karplus, J. A. McCammon, *Nat. Struct. Biol.* **2002**, *9*, 646.

[115] J. D. Durrant, J. A. McCammon, *BMC Biol.* **2011**, *9*, 71.

[116] H. A. L. Filipe, L. M. S. Loura, *Molecules* **2022**, *27*.

[117] W. F. van Gunsteren, C. Oostenbrink, *J. Chem. Inf. Model.* **2024**, *64*, 6281.

[118] J. Jung, W. Nishima, M. Daniels, G. Bascom, C. Kobayashi, A. Adedoyin, M. Wall, A. Lappala, D. Phillips, W. Fischer et al., *J. Comp. Chem.* **2019**, *40*, 1919.

[119] L. Verlet, *Phys. Rev.* **1967**, *159*, 98.

[120] D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke,A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C.Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I.Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails,R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman, *AMBER 2016*, University of California San Francisco, San Francisco, **2016**.

[121] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.

[122] Q. Ke, X. Gong, S. Liao, C. Duan, L. Li, *J. Mol. Liq.* **2022**, *365*, 120116.

[123] H. C. Andersen, *J. Chem. Phys.* **1980**, *72*, 2384.

[124] W. F. van Gunsteren, H. J. C. Berendsen, *Mol. Simul.* **1988**, *1*, 173.

[125] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684.

[126] G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 14101.

[127] S. Nosé, *J. Chem. Phys.* **1984**, *81*, 511.

[128] W. G. Hoover, *Phys. Rev. A Gen. Phys.* **1985**, *31*, 1695.

[129] M. Parrinello, A. Rahman, *J. Appl. Phys.* **1981**, *52*, 7182.

[130] E. Paquet, H. L. Viktor, *Adv. Chem.* **2018**, *2018*, 1.

[131] R. N. Barnett, U. Landman, *Phys. Rev. B Condens. Matter* **1993**, *48*, 2081.

[132] R. Car, M. Parrinello, *Phys. Rev. Lett.* **1985**, *55*, 2471.

[133] M. Born, V. Fock, *Z. Physik* **1928**, *51*, 165.

[134] D. Marx, J. Hutter, *Ab Initio Molecular Dynamics*. *Basic Theory and Advanced Methods*, Cambridge University Press, **2009**.

[135] A. H. Mazurek, Ł. Szeleszczuk, D. M. Pisklak, *Int. J. Mol. Sci.* **2021**, *22*.

[136] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.

[137] J.-D. Chai, M. Head-Gordon, *J. Chem. Phys.* **2008**, *128*, 84106.

[138] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, S. Tretiak, *Sci Data* **2020**, *7*, 134.

[139] B. D. Sellers, N. C. James, A. Gobbi, *J. Chem. Inf. Model.* **2017**, *57*, 1265.

[140] M. Rezaee, S. Ekrami, S. M. Hashemianzadeh, *Sci Rep* **2024**, *14*, 11791.

[141] J. Behler, *J. Chem. Phys.* **2011**, *134*, 74106.

[142] M. Rosenblatt, *Ann. Math. Statist.* **1956**, *27*, 832.

[143] J. R. Pliego, J. M. Riveros, *WIREs Comput. Mol. Sci.* **2020**, *10*.

[144] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999.

[145] S. Miertuš, E. Scrocco, J. Tomasi, *Chem. Phys.* **1981**, *55*, 117.

[146] A. Klamt, G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

[147] A. Klamt, *J. Phys. Chem.* **1995**, *99*, 2224.

[148] E. Cancès, B. Mennucci, J. Tomasi, *J. Chem. Phys.* **1997**, *107*, 3032.

[149] V. Barone, M. Cossi, *J. Phys. Chem. A* **1998**, *102*, 1995.

[150] R. Cammi, V. Verdolino, B. Mennucci, J. Tomasi, *Chem. Phys.* **2008**, *344*, 135.

[151] R. Cammi, *J. Comput. Chem.* **2015**, *36*, 2246.

[152] B. Chen, K. N. Houk, R. Cammi, *Chem. Eur. J.* **2022**, *28*, e202200246.

[153] M. Pagliai, R. Cammi, G. Cardini, V. Schettino, *J. Phys. Chem. A* **2016**, *120*, 5136.

[154] J. Laranjeira, R. Lang, R. Cammi, B. Chen, *J. Chem. Theory Comput.* **2026**.

[155] M. Scheurer, A. Dreuw, E. Epifanovsky, M. Head-Gordon, T. Stauch, *J. Chem. Theory Comput.* **2021**, *17*, 583.

[156] D. Kang, J. Dai, J. Yuan, *J. Chem. Phys.* **2011**, *135*, 24505.

[157] N. D. Mermin, *Phys. Rev.* **1965**, *137*, A1441-A1443.

[158] J.-P. Hansen, I. R. McDonald, *Theory of simple liquids*. *With applications to soft matter*, Elsevier, Amsterdam, **2013**.

[159] L. S. Ornstein, F. Zernike, *Proc. Acad. Sci. Amsterdam* **1914**.

[160] F. Hirata, *Molecular theory of solvation*, Kluwer Acad. Publ, Dordrecht, **2003**.

[161] L. Blum, A. J. Torruella, *J. Chem. Phys.* **1972**, *56*, 303.

[162] L. Blum, *J. Chem. Phys.* **1972**, *57*, 1862.

[163] D. Chandler, H. C. Andersen, *J. Chem. Phys.* **1972**, *57*, 1930.

[164] H. C. Andersen, D. Chandler, *J. Chem. Phys.* **1972**, *57*, 1918.

[165] T. Morita, *Prog. Theor. Phys.* **1958**, *20*, 920.

[166] D. Beglov, B. Roux, *J. Phys. Chem. B* **1997**, *101*, 7821.

[167] A. Kovalenko, F. Hirata, *Chem. Phys. Lett.* **1998**, *290*, 237.

[168] C. Hölzl, P. Kibies, S. Imoto, R. Frach, S. Suladze, R. Winter, D. Marx, D. Horinek, S. M. Kast, *J. Chem. Phys.* **2016**, *144*, 144104.

[169] S. J. Singer, D. Chandler, *Mol. Phys.* **1985**, 621.

[170] S. M. Kast, T. Kloss, *J. Chem. Phys.* **2008**, *129*, 236101.

[171] A. Kovalenko, F. Hirata, *J. Chem. Phys.* **2000**, *112*, 10391.

[172] H. A. Lorentz, *Ann. Phys.* **1881**, *248*, 127.

[173] J. Heil, S. M. Kast, *J. Chem. Phys.* **2015**, *142*, 114107.

[174] N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, S. M. Kast, *J. Comput. Aided Mol. Des.* **2016**, *30*, 1035.

[175] N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput. Aided Mol. Des.* **2018**, *32*, 1151.

[176] D. Roy, A. Kovalenko, *J. Phys. Chem. A* **2019**, *123*, 4087.

[177] N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput. Aided Mol. Des.* **2020**, *34*, 453.

[178] M. Reimann, M. Kaupp, *J. Phys. Chem. A* **2020**, *124*, 7439.

[179] R. Frach, S. M. Kast, *J. Phys. Chem. A* **2014**, *118*, 11620.

[180] S. Ten-no, F. Hirata, S. Kato, *Chem. Phys. Lett.* **1993**, *214*, 391.

[181] S. Ten-no, F. Hirata, S. Kato, *J. Chem. Phys.* **1994**, *100*, 7443.

[182] A. Kovalenko, F. Hirata, *J. Chem. Phys.* **1999**, *110*, 10095.

[183] H. Sato, A. Kovalenko, F. Hirata, *J. Chem. Phys.* **2000**, *112*, 9463.

[184] N. Minezawa, S. Kato, *J. Chem. Phys.* **2007**, *126*, 54511.

[185] S. Aono, S. Sakaki, *J. Phys. Chem. B* **2012**, *116*, 13045.

[186] D. Yokogawa, H. Sato, S. Sakaki, *J. Chem. Phys.* **2007**, *126*, 244504.

[187] D. Yokogawa, *Chem. Phys. Lett.* **2013**, *587*, 113.

[188] K. Imamura, D. Yokogawa, H. Sato, *J. Chem. Phys.* **2024**, *160*.

[189] G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, T. Ziegler, *J. Comput. Chem.* **2001**, *22*, 931.

[190] Y. Maruyama, N. Yoshida, *J. Comput. Chem.* **2024**, *45*, 1470.

[191] Y. Harano, T. Imai, A. Kovalenko, M. Kinoshita, F. Hirata, *J. Chem. Phys.* **2001**, *114*, 9506.

[192] M. Misin, M. V. Fedorov, D. S. Palmer, *J. Phys. Chem. B* **2016**, *120*, 975.

[193] M. D. Tissandier, K. A. Cowen, W. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, T. R. Tuttle, *J. Phys. Chem. A* **1998**, *102*, 7787.

[194] D. S. Palmer, A. I. Frolov, E. L. Ratkova, M. V. Fedorov, *J. Phys. Condens. Matter* **2010**, *22*, 492101.

[195] V. Sergiievskyi, G. Jeanmairet, M. Levesque, D. Borgis, *J. Chem. Phys.* **2015**, *143*, 184116.

[196] A. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. K. Giesen, P. Winget, C. J. Cramer, D. G. Truhlar, *Minnesota solvation database (MNSOL) version 2012*, **2020**.

[197] A. Ben-Naim, *J. Phys. Chem.* **1978**, *82*, 792.

[198] A. Gelman, X.-L. Meng, *Statist. Sci.* **1998**, *13*.

[199] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, *J. Chem. Phys.* **2020**, *152*, 224108.

[200] S. Gusarov, T. Ziegler, A. Kovalenko, *J. Phys. Chem. A* **2006**, *110*, 6083.

[201] F. Bloch, *Phys. Rev.* **1946**, *70*, 460.

[202] E. M. Purcell, R. V. Pound, N. Bloembergen, *Phys. Rev.* **1946**, *70*, 986.

[203] ZAVOISKY E., *J. Phys. U. S. S. R.* **1945**, *9*, 211p.

[204] R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, P. Granger, R. E. Hoffman, K. W. Zilm, *Pure Appl. Chem.* **2008**, *80*, 59.

[205] R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, R. Goodfellow, P. Granger, *Pure Appl. Chem.* **2001**, *73*, 1795.

[206] D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, B. D. Sykes, *J. Biomol. NMR* **1995**, *6*, 135.

[207] R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, R. Goodfellow, P. Granger, *Solid State Nucl. Magn. Reson.* **2002**, *22*, 458.

[208] N. J. Bunce, *J. Chem. Educ.* **1987**, *64*, 907.

[209] E. Fermi, *Z. Physik* **1930**, *60*, 320.

[210] T. D. W. Claridge, *High-resolution NMR techniques in organic chemistry*, Elsevier, Amsterdam, Boston, Heidelberg London, New York, Oxford, Paris, **2016**.

[211] N. F. Ramsey, *Phys. Rev.* **1950**, *78*, 699.

[212] N. F. Ramsey, *Phys. Rev.* **1952**, *86*, 243.

[213] N. F. Ramsey, *Phys. Rev.* **1953**, *91*, 303.

[214] M. Kaupp, M. Bühl, V. G. Malkin, *Calculation of NMR and EPR Parameters*. *Theory and applications,* 1. Aufl., Wiley-VCH, Weinheim, **2006**.

[215] W. Kutzelnigg, *Isr. J. Chem.* **1980**, *19*, 193.

[216] M. Schindler, W. Kutzelnigg, *J. Chem. Phys.* **1982**, *76*, 1919.

[217] T. A. Keith, R. F. Bader, *Chem. Phys. Lett.* **1993**, *210*, 223.

[218] T. A. Keith, R. Bader, *Chem. Phys. Lett.* **1992**, *194*, 1.

[219] K. Ruud, T. Helgaker, K. L. Bak, P. Jo/rgensen, H. J. A. Jensen, *J. Chem. Phys.* **1993**, *99*, 3847.

[220] F. London, *J. Phys. Radium* **1937**, *8*, 397.

[221] C. N. Yang, *Phys. Today* **2014**, *67*, 45.

[222] A. M. Sarotti, S. C. Pellegrinet, *J. Org. Chem.* **2009**, *74*, 7254.

[223] L. Eberlein, *A combined computational and NMR-spectroscopic approach for tautomer elucidation under extreme conditions towards investigating the robustness of genetic codes, Doctoral Dissertation*, Dortmund, **2021**.

[224] T. Pongratz, *Rism-based pressure-dependent computational spectroscopy, Doctoral Dissertation*, TU Dortmund, Dortmund, **2022**.

[225] M. R. Willcott, *J. Am. Chem. Soc.* **2009**, *131*, 13180.

[226] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1-2*, 19.

[227] J. L. F. Abascal, C. Vega, *J. Chem. Phys.* **2005**, *123*, 234505.

[228] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, W. L. Jorgensen, *Nucleic. Acids. Res.* **2017**, *45*, W331-W336.

[229] M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **2015**, *11*, 3499.

[230] W. L. Jorgensen, J. Tirado-Rives, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6665.

[231] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696.

[232] L. Kagami, A. Wilter, A. Diaz, W. Vranken, *Bioinformatics* **2023**, *39*.

[233] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577.

[234] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *J. Comput. Chem.* **1997**, *18*, 1463.

[235] S. Imoto, H. Forbert, D. Marx, *Phys. Chem. Chem. Phys.* **2015**, *17*, 24224.

[236] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652.

[237] I. Pickering, J. A. Semelak, J. Xue, A. E. Roitberg, *J. Phys. Chem. B* **2025**, *129*, 11927.

[238] J. G. Brandenburg, C. Bannwarth, A. Hansen, S. Grimme, *J. Chem. Phys.* **2018**, *148*, 64104.

[239] Y. Xue, *Personal communication*, **2023**, Gainesville.

[240] R. Zwanzig, *J Stat Phys* **1973**, *9*, 215.

[241] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.

[242] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji et al., *Gaussian 16 Rev. C.01*, Wallingford, CT, **2016**.

[243] A. Kovalenko, S. Ten-no, F. Hirata, *J. Comput. Chem.* **1999**, *20*, 928.

[244] K. M. Kast, J. Brickmann, S. M. Kast, R. S. Berry, *J. Phys. Chem. A* **2003**, *107*, 5342.

[245] Z. A. Makrodimitri, V. E. Raptis, I. G. Economou, *J. Phys. Chem. B* **2006**, *110*, 16047.

[246] L. E. Chirlian, M. M. Francl, *J. Comput. Chem.* **1987**, *8*, 894.

[247] W. B. Floriano, M. A. C. Nascimento, *Braz. J. Phys.* **2004**, *34*, 38.

[248] V. A. Tran, M. Teucher, L. Galazzo, B. Sharma, T. Pongratz, S. M. Kast, D. Marx, E. Bordignon, A. Schnegg, F. Neese, *J. Phys. Chem. A* **2023**, *127*, 6447.

[249] M. Saitow, U. Becker, C. Riplinger, E. F. Valeev, F. Neese, *J. Chem. Phys.* **2017**, *146*, 164105.

[250] C. Riplinger, F. Neese, *J. Chem. Phys.* **2013**, *138*, 34106.

[251] S. Grimme, A. Hansen, S. Ehlert, J.-M. Mewes, *J. Chem. Phys.* **2021**, *154*, 64103.

[252] J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graphics Modell.* **2006**, *25*, 247.

[253] J. G. Kirkwood, F. P. Buff, *J. Chem. Phys.* **1951**, *19*, 774.

[254] T. Poggio, F. Girosi, *Science* **1990**, *247*, 978.

[255] K. K. Huddleston, *Personal communication*, **2023**, Gainesville.

[256] X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith, A. E. Roitberg, *J. Chem. Inf. Model.* **2020**, *60*, 3408.

[257] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, **2014**.

[258] I. Loshchilov, F. Hutter, *Decoupled Weight Decay Regularization*, **2017**.

[259] D. Sebastiani, M. Parrinello, *ChemPhysChem* **2002**, *3*, 675.

[260] M. P. Williamson, *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *73*, 1.

[261] P. Dobrev, S. P. B. Vemulapalli, N. Nath, C. Griesinger, H. Grubmüller, *J. Chem. Theory Comput.* **2020**, *16*, 2561.

[262] L. Eisel, *Multiscale approximations of integral equation-based solvation models, Doctoral Dissertation*, Dortmund, **2025**.

[263] H. R. Kalbitzer, *Personal communication.*, 2024.

[264] S. Maste, *Hybride implizite und explizite Solvatationsansätze zur Berechnung von NMR-Parametern, Master's thesis*, Dortmund, **2020**.

[265] S. Maste, B. Sharma, T. Pongratz, B. Grabe, W. Hiller, M. B. Erlach, W. Kremer, H. R. Kalbitzer, D. Marx, S. M. Kast, *Phys. Chem. Chem. Phys.* **2024**, 6386.

[266] W. M. Litchman, M. Alei, A. E. Florin, *J. Am. Chem. Soc.* **1969**, *91*, 6574.

[267] L. Galazzo, S. Maste, B. Sharma, A. van Tran, T. Pongratz, M. Teucher, D. Marx, F. Neese, S. M. Kast, E. Bordignon, *J. Phys. Chem. Lett.* **2025**, 8141.

[268] S. Lee, D. P. Ames, *J. Chem. Phys.* **1984**, *81*, 4206.

[269] M. F. Ottaviani, G. Martini, L. Nuti, *Magn. Reson. Chem.* **1987**, *25*, 897.

[270] A. Savitsky, M. Plato, K. Möbius, *Appl Magn Reson* **2010**, *37*, 415.

[271] D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman et al., *The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models*, **2025**.

[272] D. P. Kovács, I. Batatia, E. S. Arany, G. Csányi, *J. Chem. Phys.* **2023**, *159*.

[273] M. Windhofer, *Leistungsfähigkeit des r2SCAN-3c-DFT-Funktionals unter EC-RISM-Solvatationsbedingungen*, Dortmund, **2022**.

[274] O. Wahl, T. Sander, *J. Chem. Inf. Model.* **2020**, *60*, 1085.

[275] M. Strobl, *Personal communication*, **2025**, Dortmund.

[276] M. Wieder, J. Fass, J. D. Chodera, *Chem. Sci.* **2021**, *12*, 11364.

[277] X. Pan, F. Zhao, Y. Zhang, X. Wang, X. Xiao, J. Z. H. Zhang, C. Ji, *J. Chem. Inf. Model.* **2023**, *63*, 1833.

Eidesstattliche Versicherung (Affidavit)

Name, Vorname
(Surname, first name)

Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum
(Place, date)

Unterschrift
(Signature)