A combined computational and NMR-spectroscopic approach for tautomer elucidation under extreme conditions towards investigating the robustness of genetic codes

Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) vorgelegt von Lukas J. Eberlein

Diese Dissertation wurde im Zeitraum vom 01.08.2016 bis zum 22.01.2021 an der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt

Erstgutachter: Prof. Dr. Stefan M. Kast
Zweitgutachter: Prof. Dr. Roland Winter

Acknowledgements

Since all the people mentioned are native speakers, the following part of the section is written in german.

Zuerst möchte ich mich an dieser Stelle natürlich bei Stefan Kast bedanken. Nicht nur für die Möglichkeit diese Arbeit in deiner Arbeitsgruppe anzufertigen, deine wissenschaftliche Hilfestellung und die vielen Dinge die ich gelernt habe, sondern auch dafür, dass man jederzeit mit allen Problemen zu dir kommen kann. Ich habe mich in deiner Arbeitsgruppe immer sehr wohl gefühlt und werde es bestimmt auch in Zukunft. Ich hatte hier eine schöne Zeit und werde diese Jahre immer positiv in Erinnerung behalten.

Bei Roland Winter bedanke ich mich für die Übernahme des Zweitgutachtens, aber auch für alle weiteren Dinge die sie für mich getan haben. Sie haben mich von der Betreuung meiner Bachelorarbeit, über die Hochdruck-Forschergruppe bis jetzt in meinem bisherigen akademischen Leben begleitet.

Bei der gesamten Arbeitsgruppe Kast möchte ich mich für die gute Atmosphäre und Zusammenarbeit bedanken. Vor allem bei Patrick, der mich schon in meiner Masterarbeit betreut hat, und Tim mit denen ich unter anderem in der Hochdruck-Gruppe zusammengearbeitet habe und viele schöne Erlebnisse hatte (unter anderem Japan, Perugia und natürlich die abenteuerliche Rückfahrt aus Regensburg). Ich hoffe wir bleiben auch weiter in Kontakt nachdem die Zeit an der Uni vorüber ist. Das gilt natürlich auch für die anderen Kollegen (hoffentlich vergesse ich niemanden) Maddin, Jochen, Lars, Leo, Roland, Daniel, Chris, Stefan, Fabian, Florian, Yannic und Lenny. Bei Roland Sigel und Silke Johannsen möchte ich mich für die Möglichkeit mein „RESOLV Internship" in Zürich verbringen zu können, die Hilfe bei den NMR-Experimenten und generell die schöne Zeit bei euch bedanken. Bei Anneliese Ahlke, Andrea Jeworrek und Julia Nowack möchte ich mich für die viele Hilfe in organisatorischen Dingen an der Uni, in der Forschergruppe und bei RESOLV bedanken. Egal mit welchen absurden Fragen ich zu euch kam, ihr wusstet immer weiter.

Natürlich möchte ich mich auch noch bei meinen Freunden, um ein paar namentlich zu erwähnen, Tim, Sara, Nico, Chrissy, Christian, Jonny, Herbie, Felix, Mathis, Peder, und die Gang, meinen Mitbewohnerinnen Theresa und Linda und meiner Familie, Peter, Astrid, Daria, Yako, Merle, Felix, Moritz, Judith, Lukas und Sigrun für die ganze Unterstützung in allen Lebenslagen bedanken. Und vor allem bei Julia, einfach dafür das du da bist.

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)

Contents

Acknowledgements ii

Eidesstattliche Versicherung iii

I. Introduction 1

Abstract 1

Zusammenfassung 3

Publications 5

1.1 What is tautomerism? 6

1.2 Importance of tautomerism in biological systems 6

1.3 Experimental insights into tautomerism 7

1.4 Computational insights into tautomerism 9

1.5 Tautomerism of 2-pyridone / 2-hydroxypyridine 11

1.6 Tautomerism in nucleic acids and their building blocks 12

1.7 Non-natural nucleic acids – the Hachimoji code 14

1.8 Thesis outline 17

II. Theory 18

2.1 Basics of the nuclear magnetic resonance spectroscopy 18

2.2 Basics of quantum chemical calculations 20

2.2.1 The Hartree-Fock method 20

2.2.2 Møller-Plesset perturbation theory 22

2.2.3 Coupled Cluster Theory 23

2.2.4 Density functional theory in quantum chemistry 26

2.2.5 Calculation of nuclear magnetic resonance parameters 27

2.2.6 Basis sets 30

2.3 Classical mechanics 32

2.3.1 Molecular mechanics 32

2.3.2 Molecular dynamics simulations 33

2.4 Solvation models 34

2.4.1 Implicit solvation models / Thermodynamic integration 35

2.4.2 Dielectric continuum models 38

2.4.3 Statistical solvation models / Density functional theory 40

2.4.4 The reference interaction site model 44

2.5 The Embedded-Cluster Reference Interaction Site Model 48

2.5.1 Basics 48

2.5.2 Empirical correction terms for EC-RISM – the partial molar volume correction 51

2.5.3 The partial molar volume correction for extreme conditions 52

2.6 Calculation of conformational, tautomer and ionization state fractions and dissociation constants. 55

III. Methods 59

3.1 General computational details 59

3.1.1 Geometry optimizations and single point calculations 59

3.1.2 EC-RISM calculations 59

3.1.3 Generation of solvent susceptibilities for various temperatures 60

3.1.4 TI calculations 62

3.2 Computational details of the specific chapters 63

3.2.1 EC-RISM for temperature variations: Parametrization and benchmarking 63

3.2.2 Calibrating tautomer predictions: the SAMPL2 datasets 63

3.2.3 Probing the sensitivity of chemical shifts for tautomer predictions - histamine 64

3.2.4 Calibrating the prediction of NMR chemical shifts 68

3.2.5 Tautomerism of nucleobases: Natural nucleobases, Hachimoji bases and derivatives 69

3.2.6 Nucleotides at high hydrostatic pressure: NMR of adenosine monophosphate 69

3.2.7 Nucleotides at various temperatures 69

IV. Results 74

4.1 EC-RISM for temperature variations: Parametrization and benchmarking 74

4.2 Calibrating tautomer predictions: the SAMPL2 dataset 80

4.3 Probing the sensitivity of chemical shifts for tautomer predictions - histamine 90

4.3.1 Reparametrization of GAFF parameters 93

4.3.2 Calculation of histamine tautomerism 96

4.3.3 Calculation of histamine NMR chemical shifts 98

4.3.4 Extraction of histamine tautomer fractions from NMR chemical shifts 104

4.4 Calibrating the prediction of NMR chemical shifts 109

4.4.1 High hydrostatic pressure conditions 109

4.4.2 Variation of temperature conditions 117

4.5 Tautomerism of nucleobases: Natural nucleobases, Hachimoji bases and derivatives 127

4.5.1 Investigated tautomers 127

4.5.2 Ambient conditions 132

4.5.3 High hydrostatic pressure conditions 139

4.5.4 High temperature conditions 144

4.6 Nucleotides at high hydrostatic pressure: NMR of adenosine monophosphate 151

4.7 Nucleotides at high temperatures 156

4.7.1 Experimental results 156

4.7.2 Nucleosides 160

4.7.3 Nucleotides 167

V. Summary 185

References 189

Appendix 201

I. Introduction

Abstract

Nucleic acids are crucial for live. The central dogma of molecular biology, stated by Francis Crick, is that genetic information is stored in one nucleic acid (DNA), is transferred to another one (RNA), and is then translated into the amino acid sequence of proteins. The reverse, however, the transfer from the protein sequence back to the nucleic acid, is impossible.[1] In this work, the stability of nucleic acids and therefore the information needed for life is studied with respect to the tautomeric stability of their components, nucleobases and nucleotides. The assumed origin of life is the deep sea, where oceanic black and white smokers offer conditions that make them likely to be the place where the first building blocks of life were formed.[2] At these deep sea vents, high temperatures and pressures occur, making it important to understand the influence of these extreme environmental conditions on tautomer stability of nucleic acid building blocks, which is related to the mutation rate in nucleic acids, since the tautomeric state of a nucleobase determines the protonation pattern presented to a potential hydrogen bonding partner in order to form a base pair. A switch in the tautomeric state leads to a different protonation pattern and therefore different possible hydrogen bonds formed with partners, which can directly be related to the incorporation of a wrong base in the opposing strand; a mutation occurs. Elucidation of this is not possible by simply calculating the quantum chemical (QC) energetics of the respective species, but also requires consideration of the mutual polarization of both solvent and solute at given environmental conditions. This can be achieved using the embedded cluster reference interaction site model (EC-RISM) and suitable correction terms for extreme conditions.

Once the first living systems have developed, it is crucial that they were able to undergo Darwinian evolution to develop into more complex life forms. This gives rise to the question how robust and universal our genetic code is in the universe. Is our genetic code the only one suitable for this task, or are other, alternative genetic codes, like the Hachimoji code[3] which is an 8-letter extension to the natural genetic code, also capable of undergoing Darwinian evolution, maybe even better at extreme environmental conditions? To answer these questions, the theoretical calculations done in this work are supported by complementary experiments using

the nuclear magnetic resonance (NMR) spectroscopy. For comparison with computed spectra, the calculation of suitable reference substances, especially for extreme conditions, is needed in order to obtain a reliable workflow for NMR predictions. In this work, the reference substance dimethyl-silapentane-sulfonate (DSS) is used, which, like nucleotides, has a lot of conformational degrees of freedom, needing a computational workflow resulting in a sufficiently sampled conformational ensemble. The reliability of computational tautomer predictions is verified not only by NMR experiments, but also benchmarking with known databases is done. Finally, tautomers and ionization states are investigated for the small neurotransmitter histamine, which has a lot of conformations. Also, an approach to optimize some of the force field (FF) parameters in the general amber force field (GAFF) used for EC-RISM calculations is developed.

With the workflows and results obtained during this work, it is possible to calculate NMR chemical shifts over a broad range of temperatures and pressures using the EC-RISM solvation model. These chemical shifts as well as (temperature dependent) energetics can be used to predict tautomer ratios with high accuracy. Using the model system histamine, it is shown that EC-RISM can be used to improve FF parameters for the prediction of pK_a-values in aqueous solution. The benchmarking, the improvements developed during this work, and NMR spectroscopic experiments are successfully applied for the investigation of nucleic acid building blocks. The natural nucleobases, nucleosides and nucleotides are tautomer stable over a broad range of pressures and temperatures, while the extended Hachimoji code needs some further improvement.

Zusammenfassung

Nucleinsäuren sind elementare Bausteine des Lebens. Das zentrale Dogma der Molekularbiologie, welches von Francis Crick formuliert wurde, besagt, dass die Erbinformationen in einer Nucleinsäure (DNS) gespeichert sind, von dort in eine andere überführt werden (RNS) und anschließend in die Aminosäuresequenz der Proteine übersetzt werden. Hingegen ist der umgekehrte Prozess, der Transfer der Proteinsequenz in eine Nucleinsäure, unmöglich. [1] In dieser Arbeit wird die Stabilität von Nucleinsäuren, und damit der lebenswichtigen Erbinformationen, im Hinblick auf die Tautomer Stabilität ihrer Komponenten, der Nucleobasen und Nucleotide, studiert. Als Ursprung des Lebens wird die Tiefsee vermutet, dort bieten schwarze und weiße Raucher Umgebungsbedingungen die es wahrscheinlich machen, dass dort die ersten Bausteine des Lebens geformt wurden.[2] An diesen Tiefseequellen herrschen hohe Temperaturen und Drücke, daher ist es wichtig, den Einfluss dieser extremen Umweltbedingungen auf die Stabilität der Nucleinsäurebausteine zu überprüfen. Die vorliegenden Tautomere dieser Bausteine bestimmen das Protonierungsmuster, welches potentiellen Wasserstoffbrücken-Bindungspartnern präsentiert wird um ein Basenpaar zu bilden, daher kann die Tautomer Stabilität direkt mit der Mutationsrate in Nucleinsäuren korreliert werden. Ein Wechsel des Tautomers kann zu einem Wechsel im Protonierungsmuster und damit zu anderen Möglichkeiten Wasserstoffbrücken auszubilden führen, was wiederum zu einem anderen Bindungspartner führt, der anschließend in den Gegenstrang eingebaut wird; dies ist eine Punktmutation. Um diese Stabilität zu überprüfen, reichen einfache quantenchemische (QC) Energieberechnungen nicht aus, da die gegenseitige Polarisation von gelöstem Teilchen und Lösungsmittel bei den gegebenen Umgebungsbedingungen berücksichtig werden muss. Dies kann jedoch mit dem „embedded cluster reference interaction site model“ (EC-RISM) und für die Extrembedingungen geeigneten Korrekturtermen erreicht werden.

Nachdem sich die ersten lebenden Systeme entwickelt hatten, war es wichtig, dass sie in der biologischen Evolution unterliegen, um sich zu komplexeren Lebensformen zu entwickeln. Daraus ergibt sich die Frage nach der Stabilität und der Allgemeinheit unseres genetischen Codes im Universum. Ist unser genetischer Code der einzige, der in der Lage ist, der Evolution unterliegende Systeme zu entwickeln oder sind andere Codes, wie der Hachimoji-Code,[3] der eine Erweiterung unseres genetischen Codes auf acht Buchstaben ist, ebenfalls dazu in der Lage? Eventuell sogar besser bei extremen Umgebungsbedingungen? Um diese Fragen zu beantworten, sind in dieser Arbeit nicht nur theoretische Berechnungen durchgeführt worden, sondern auch komplementäre Kernspinresonanzspektroskopie (NMR) Experimente durchgeführt worden.

Um experimentelle und berechnete Spektren vergleichen zu können, ist es nötig geeignete Referenzsubstanzen zu betrachten und deren Eigenschaften zu berechnen, besonders unter extremen Umgebungsbedingungen. In dieser Arbeit wird der Fokus auf die Referenzsubstanz 2,2-Dimethyl-2-silapentan-5-sulfonsäure (DSS), eine Substanz die ähnlich wie Nucleotide viele konformationelle Freiheitsgrade hat, gelegt. Um diese Freiheitsgrade in Rechnungen zu berücksichtigen, wird ein Arbeitsablauf benötigt der ein möglichst realistisches konformationelles Ensemble erzeugt. Um die Zuverlässigkeit von berechneten Tautomerenverhältnissen zu überprüfen wurden nicht nur NMR Experimente durchgeführt, sondern auch „Benchmark“ Rechnungen an Datenbanken mit experimentellen Ergebnissen durchgeführt. Außerdem wurden die Tautomerenverhältnisse und Ionisierungsstufen des Neutrotransmitters Histamine, eines kleinen Moleküls mit vielen konformationellen Freiheitsgraden, untersucht. Dabei wurde ein Arbeitsablauf um Kraftfeldparameter (FF Parameter) des „general amber force field“ (GAFF) welches für EC-RISM Rechnungen benötigt wird, entwickelt.

Mit den in dieser Arbeit entwickelten Arbeitsabläufen und den damit erhaltenen Ergebnissen ist es möglich chemische Verschiebungen mithilfe von EC-RISM für eine weite Spanne von Temperaturen und Drücken zu berechnen. Diese chemischen Verschiebungen und die temperaturabhängigen Energien von Molekülen können genutzt werden um Tautomerenverhältnisse mit hoher Genauigkeit zu berechnen. Am Beispiel des Histamins wird gezeigt, dass mithilfe von EC-RISM FF Parameter optimiert werden können um die Berechnung von pK_a-Werten in Lösung zu verbessern. Das „Benchmarking“ und die Weiterentwicklungen die in dieser Arbeit vorgestellt werden, werden, zusammen mit den NMR Experimenten, erfolgreich für die Untersuchung der Nucleinsäurebausteine eingesetzt. Als Ergebnis wird erhalten, dass die natürlichen Bausteine über eine weiten Temperatur- und Druckbereich tautomerenstabil sind, während der Hachimoji-Code dafür weitere Verbesserungen benötigt.

Publications

Parts of this work are already available in the following publications under participation of the author:

1: N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, "The SAMPL6 challenge on predicting aqueous pK_a values from EC-RISM theory", *J. Comput.-Aided Mol. Des*. 32, 1151 (2018)[4]

2: N. Tielker, L. Eberlein, C. Chodun, S. Güssregen, S. M. Kast, "pK_a calculations for tautomerizable and conformationally flexible molecules: partition function vs. state transition approach", *J. Mol. Model*. 25, 139 (2019)[5]

3: C. E. Munte, M. Karl, W. Kauter, L. Eberlein, T.-V. Pham, M. Beck Erlach, S. M.Kast, W. Kremer, H. R.Kalbitzer, "High pressure response of ^{1}H NMR chemical shifts of purine nucleotides", *Biophys. Chem*. 254, 106261 (2019)[6]

4: T. Pongratz, P. Kibies, L. Eberlein, N. Tielker, C. Hölzl, S. Imoto, M. Beck Erlach, S. Kurrmann, P. H. Schummel, M. Hofmann, O. Reiser, R. Winter, W. Kremer, H. R. Kalbitzer, D. Marx, D. Horinek, S. M. Kast, "Pressure-dependent electronic structure calculations using integral equation-based solvation models", *Biophys. Chem.* 257, 106258 (2020)[7]

5: N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, S. M. Kast, "The SAMPL6 challenge on predicting octanol-water partition coefficients from EC-RISM theory", *J. Comput.-Aided Mol. Des*. 34, 453 (2020)[8]

6: L. Eberlein, F. R. Beierlein, N. J. R. van Eikema Hommes, A. Radadiya, J. Heil, S. A. Benner, T. Clark, S. M. Kast, N. G. J. Richards, "Tautomeric equilibria of nucleobases in the Hachimoji expanded genetic alphabet", *J. Chem. Theory Comput*. 16, 4, 2766 (2020)[9]

7: N. Tielker, L. Eberlein, G. Hessler, K. F. Schmidt, S. Güssregen, S. M. Kast, "Quantum-mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges?", *J. Comput.-Aided Mol. Des*. https://doi.org/10.1007/s10822-020-00347-5 (2020)[10]

1.1 What is tautomerism?

The tautomeric effect is one major focus of this work; the term tautomerism describes the migration of functional groups or, more often, atoms within the same organic molecule. The term was introduced in 1885 by Conrad Laar for the endpoints of an intramolecular vibration[11,12] although today's definition of tautomerism is more in line with the historical alternative concept of protomerism.[13] This term prevailed and is now used for a wide range of readily interconvertible isomers where the migrating group becomes a nucleo- or electrofuge during the isomerization reaction.[14] Most of the time, the migrating atom is a hydrogen atom or a proton, especially in aqueous conditions where the solvent molecules assist the tautomerization process; this is called prototropic tautomerism and is investigated in the context of this work.

Although tautomerism is an isomeric effect,[14] it is different from typical isomeric forms such as enantiomers or *cis-*/*trans* isomers. In these forms the transformations are usually difficult, which allows isolation of the respective states. In contrast to these, the tautomeric transitions are usually much faster, making it hard to detect them separately; tautomers form a dynamic equilibrium.

1.2 Importance of tautomerism in biological systems

The small energetic differences and fast transitions lead to experimental difficulties in the analysis of tautomers, yet simultaneously allow useful applications. Tautomers can have a completely different structure, with different hydrogen bonding patterns, and thus different chemical and biological properties. In biological systems, the processes taking place have to be controlled by fine switching, which is possible due to tautomerism. Enzymes called tautomerases are catalyzing tautomerization reactions in cells and play important roles in biochemical pathways, like the dopachrome tautomerase (DCT) in the synthesis of the melanin pigment.[15,16] Another famous case in biology is the tautomerism of nucleobases, which is discussed in detail in chapter 1.6. An example system that is currently studied in the Kast group is the tautomerism in the Rizin-A/Pteroic acid-complex for which the pH-dependent tautomer populations drastically change upon ligand binding,[17,18] but there are also a multitude of artificial switchable systems involving tautomers.[19,20] Tautomerism is also an important aspect to consider during drug development, because the molecules have to maintain functionality under physiological conditions.[21] In biological systems like living cells, tautomers are influenced by a lot of environmental properties, like the solvent polarity (usually polar solvents like water favor the more polar tautomer, in contrast to for example apolar membranes), the pH value of the medium (due to the possibility to switch tautomerism on/off by changing the protonation state and a wide range of pH values in the body),[21] pressure, macromolecular crowding and, most importantly, temperature. High temperature stabilizes the least stable tautomer due to the relationship between the equilibrium constant and the standard free enthalpy described by the well-known formula $\Delta_{\mathrm{r}}G^\circ = -RT\ln K$, therefore the temperature dependence of tautomerism is an important aspect of this work.[21]

1.3 Experimental insights into tautomerism

Experimentally, there are a lot of strategies to get insight into tautomerization phenomena. A method allowing the direct observation of a tautomerizing molecule on a surface at low temperature is the scanning tunneling microscopy (STM) with which the tautomerisation of tetraphenyl-porphyrin could already be clarified;[22] the method is described in detail in Ref. 23. Another approach on a slower timescale and requiring difficult sample preparations is crystallography,[24] but to get insight in dynamic tautomerism, especially in solution, faster timescales are needed. A suitable method in the gas-phase is the vacuum ultraviolet spectroscopy (VUV).[25] In solution, flash photolysis[26] can be used, but here ionization of the molecule is needed. A selection of experimental methods like stationary and time resolved fluorescence spectroscopy, femtosecond pump-pump spectroscopy, linear solvation energy relationships (LSER), and the basicity method are discussed in detail in Ref. 27. Two of the most used experimental techniques for tautomer elucidation, the ultraviolet/visible (UV/VIS)[28] and nuclear magnetic resonance (NMR) spectroscopy will be discussed here in detail.[29]

The main advantage of UV/VIS spectroscopy is the timescale. Electron excitation is faster than the proton transfers in tautomers, allowing for their measurement as individual species even if they are physically inseparable. The method is inexpensive, flexible, and the use of different solvent and environmental conditions is possible which is advantageous as well. The problem is the overlap of the spectra of the respective tautomers, therefore the determination of the tautomer populations is difficult as long as the spectra of the individual species are unknown. This problem can be solved in different ways, for instance through the use of fixed model compounds, where one of the tautomerizing protons is exchanged with a methyl group, or by computation of the individual spectra using quantum chemical methods or advanced chemometric approaches, allowing for quantitative results.[28,30]

NMR spectroscopy in general is one of the most important methods for structure elucidation due to the broad range of applications. Experiments can be performed in the gas phase, the liquid and solid state, it allows the investigation of small molecules as well as larger systems such as polymers or proteins over a wide range of temperatures (-180 to +200 °C) and pressures (regularly up to approximately 3 kbar, with up to 9 kbar in extreme cases).[31,32,33,34] Various NMR parameters are experimentally accessible, such as chemical shift, the nuclear Overhauser effect (NOE), and spin-spin coupling constants, which are all suited for structure elucidation. These parameters contain several pieces of information about the molecules under investigation. The NOE allows very precise determination of atomic distances,[35] ^3J couplings of dihedral angles[36,37] and ^1J couplings of atom-distances.[38] Multiple methods are based on the angle dependence of coupling constants and allow the determination of relative[39] and absolute[40,41] configurations.[42] These methods are not suited for tautomer elucidation, due to the fast proton transfer, especially in protic solvents; here, the measurement of the exchanging protons is often impossible even at low temperatures, and ensemble averages are usually obtained due to the slow time scale of NMR.[29,43] Like in UV/VIS spectroscopy, the problem of unknown properties of the individual tautomers occurs. Therefore, similar approaches are used to overcome this; one is the use of blocked derivatives of all tautomers (and calculation of the fractions of these tautomers from the spectrum of the unblocked species, which is done by methylation for the neurotransmitter histamine[44]). Alternatively, model compounds which are known to exist in only one tautomeric form can be used, as well as solid state properties for one of the tautomers (due to the fact that there more often only one tautomer exists in the solid state).[29] Other approaches are the deconvolution of the spectra[45] and the use of theoretically calculated properties (for instance from GIAO[46] or IGLO[47] calculations; this kind of calculations will play an important role in this work). Although there are several further problems, like determining the correct ionization states and the correct underlying conformational ensemble,[48] there are several examples for successful tautomer elucidations by NMR (mostly combined with QC calculations). For instance, studies about NH-benzimidazoles, investigated at liquid and solid state.[49] The main tautomers of 2-heteroaryl-substituted quinoxalines,[50] pyrazoline derivatives,[51] benzopyrano[3,4-d]imidazol-4(3H)-ones,[52] and 1-methyl-2-phenacylbenzimidazoles[53] could be determined this way. Besides, multiple structures with a possible O-H···O enol-enol tautomerism,[54] all monosubstituted methylimidazoles,[55] and numerous Schiff bases (e.g. Ref. 56) are investigated in the literature. Even if, as illustrated, NMR spectroscopy can be used to solve many problems and clarify tautomer relationships, there are limitations: For instance, if a ligand is bound to a protein, it is still possible to get insight into this binding process,[57] but the determination of chemical shifts can be very difficult due to the signal overlap, making it impossible to unambiguously identify the amino acids involved in the binding-process.[58]

1.4 Computational insights into tautomerism

Spectroscopic methods like UV/Vis- and NMR spectroscopy are powerful techniques, helping to solve different problems like tautomer elucidation. These methods can be and are nowadays supported by computational methods. For example, combined approaches were used in the publications presented in the previous section. Calculations can not only support the experimental approaches, they can even give further insight due to the possibility to calculate energetics as well as spectroscopic parameters for each conformation, tautomer and ionization state while the experiment mostly delivers ensemble information. Thus, there is mutual support of both methods. For the description of tautomerism, the proton transfer is important. Although the solvation free energetics of each tautomer can be calculated with molecular mechanics (MM) as well as with QC calculations, in classical MM approaches the minimum of the potential energy is arbitraty, which means that relative tautomer energies cannot be calculated. Also, proton transfers and therefore tautomer transitions are not obtainable. Newer techniques like proton transfer MM (additional potentials, which are assigning the proton to a respective donor or acceptor group based on geometry constraints, QM/MM approaches or reactive FFs are important techniques)[59,60] try to solve this problem and are reviewed in Ref. 61, but QC is still the method of choice for tautomer calculations. Quantum chemical approaches are very diverse with a plethora of methods: wave function methods like Hartree-Fock[62] (HF), the Møller-Plesset perturbation theory[63] (MP2), coupled-cluster theory[64,65,66] (in this work used with singles and doubles excitations with perturbed triples CCSD(T)), the quantum chemical density functional theory[67,68] (DFT) with a large variety of functionals from pure functionals such as PBE,[69,70] hybrid functionals like B3LYP[71,72,73] (the one used in this work) up to double-hybrid functionals like B2PLYP,[74] semi-empirical methods, for instance AM1[75] and PM6[76], and even *ab initio* simulations.[77,78] All these techniques have their individual strengths and weaknesses: for example, DFT, while being fast, needs empirical dispersion corrections for some calculations.[79,80,81] In this work, they are used to calculate the energetics and spectroscopic parameters of tautomers; thus multiple important aspects have to be considered: the underlying conformational ensembles (which can be calculated for example with MM simulations, as done for the epidermal growth factor receptor (EGFR) inhibitor WZ4002[82]), and the solvent model, which is crucial for the calculation of energetics and NMR parameters.[83]

The molecule geometry as well as the conformational ensemble often differ strongly between gas- and solution phase. Because of that, geometry optimizations have to be done with a model representing the respective state, such as the polarizable continuum model (PCM).[84] Consideration of these aspects has a strong influence on the computed NMR parameters.[85,86] They also depend on intermolecular interactions such as hydrogen bonds to other solute/solvent molecules which are often neglected in the gas phase or using continuum solvation.[87] For this work, the solvation model of choice for the calculation of Gibbs free energies and NMR parameters is the embedded-cluster reference interaction site model (EC-RISM).[88,89] In tautomerism, the energetic differences that have to be dealt with are often very small. It is possible to calculate these small differences correctly, but the "correct" computational method has to be chosen because computational results strongly differ depending on the chosen level of theory, basis set and solvation model. Thus, adequate benchmarking is needed, which has to be done with respect to experimental results. Besides, the computation of spectroscopic parameters can be helpful for validating the computational approach. A dataset containing experimental energetics of tautomers is available from the "statistical assessment of the modeling of proteins and ligands" challenge part two (SAMPL2), which took place in 2010. The Kast group participated with the EC-RISM solvation model resulting in the lowest overall root mean square error (RMSE) between computational prediction and experiment.[90,91] This challenge was designed in a way that the organizers provided the dataset to the participants divided into three subsets: a so-called "obscure" dataset in which only the tautomer pairs are known and the experimental data are withheld by the organizers, resulting in a blind prediction challenge; the "explanatory" dataset with unexpected experimental data which are also provided to the participants, looking for an explanation for these unexpected results; and the "investigatory" dataset for which no experimental data was available, to test the consensus of the participants.[90] This dataset is well suited for benchmarking computational approaches and will play an important role in this work.

1.5 Tautomerism of 2-pyridone / 2-hydroxypyridine

If there is one experimentally and theoretically well studied compound with respect to tautomerism, it is 2-pyridone respectively 2-hydroxypyridine. This system undergoes a lactim-lactam tautomerism, which is shown in Figure 1. Suitable ways to study this tautomerism are infrared and microwave spectroscopy, which can be done in the gas-phase, solid state and liquid state. In the gas-phase, the way to distinguish between tautomers are the OH bands of hydroxypyridine (or the C=O bands of the 2-pyridone, respectively), and a proper assignment using *ab initio* calculations. The tautomer ratio in the gas-phase is 3:1 in favor of the hydroxy-form, with at least a fraction of 95% of the Z-isomer.[92,93]

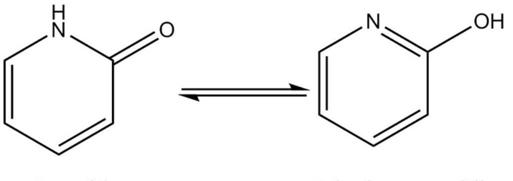


Figure 1: Tautomerism of 2-pyridone and 2-hydroxy-pyridine, a well-known tautomeric system, which is extensively described in the literature.

In the solid state, X-ray crystallographic studies are also possible, from which the relative position of the hydrogen with respect to the oxygen and nitrogen can be determined. The main form in the solid state is the 2-hydroxpyridone, where the molecules are linked to puckered chains via NH to O hydrogen bonds. The formation of dimers, which occurs in solution and is shown in Figure 2, cannot be observed.[94,95] The situation in solution is more complicated, but complementary to the experimental methods described above, also NMR measurements are well suited to investigate the system in solution. The energetic differences seem to be quite small and dependent on solvent polarity. Generally, polar media favor tautomers with larger dipole moments, therefore the 2-pyridone form is favored in media like water while 2-hydroxypyridine is favored in apolar media like cyclohexane.[96,97] Additionally, there is a dimerization observed for the system in solution. In polar and protic solvents like water, there are a lot of solute-solvent interactions and hydrogen bonds, so that the monomer is the dominant species, whereas in apolar media the dimer is preferred due to the hydrophobic effect.[98,99]

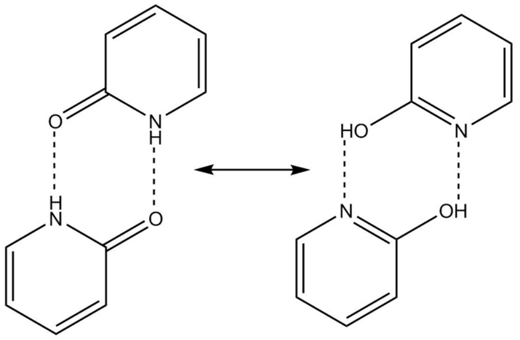


Figure 2: Two possible dimers of 2-pyridone and 2-hydroxypyridine in solution. Apolar solvents seem to prefer dimerization.

This system is part of the SAMPL2 dataset for which the experimental value for the tautomerization energy in solution is given with -4.8 kcal/mol for 2-pyridone with respect to 2-hydroxypyridine. The result of the original SAMPL2 submission of the Kast group was -7.73 kcal/mol,[91] a result that could be improved to -4.52 kcal/mol[10] with the developments during this work. This final result is in a similar order of magnitude as other QC based submissions from the SAMPL2 challenge (-4.01 kcal/mol).[100]

1.6 Tautomerism in nucleic acids and their building blocks

The central nucleic acid, storing the genetic information of higher life forms, is the DNA. The double helical structure of DNA was examined by James Watson and Francis Crick in the 1950's. There is an anecdote stating that the theoretical chemist Jerry Donahue, sharing the office with them, gave them the hint that the most probable tautomers of the nucleobases may be wrong in a lot of textbooks during this time. After using the correct tautomers, they supposedly quickly finished their work, resulting in the DNA model still valid today.[24] This shows the importance of tautomerism in biology, since only the "correct" tautomer is able to form the "correct" base pairs, a pair of cytosine and guanine, stabilized by three hydrogen bonds, and a pair of adenine and thymine (uracil in RNA), stabilized by two hydrogen bonds. These well-known Watson-Crick pairs are shown in Figure 3. Deviations from this scheme can also be observed. Especially if the strict arrangement in the double helix is abandoned, it is possible that rare tautomers occur. The A-T pair seems to adopt rare tautomers at the end of G-C rich sequences. The hints that this sequence stabilizes the rare tautomer can help to explain the sequence dependence of mutations during the DNA replication.[101] In DNA polymerases, T-G and C-A mismatches have been reported in the context of ionization events.[102]

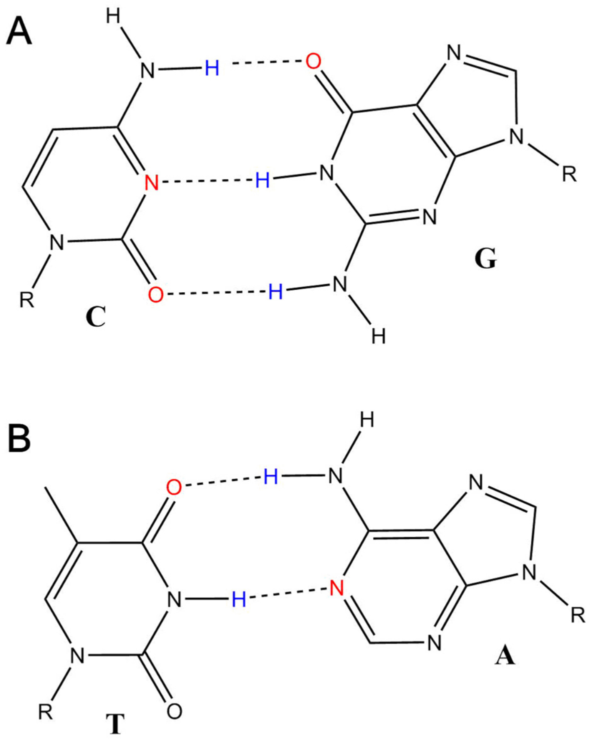


Figure 3: Natural Watson-Crick base pairs. The cytosine-guanine pair with three hydrogen bonds (A) and the thymine-adenine pair with two hydrogen bonds (B). Hydrogen bond donors are marked in blue, acceptors in red. The R denotes the ribose (RNA) or desoxyribose (DNA) bound to the base. In RNA, thymine is replaced by uracil.

Both the presence of rare tautomers during replication (exemplary mispairings are shown in Figure 4) and the formation of so-called wobble pairs, caused by unfavorable stereochemistry of the base pair, can lead to substitution mutagenesis.[103] Especially in RNA biochemistry, where a lot of catalytic processes under participation of nucleic acids occur, like self-cleaving during splicing events or riboswitches, tautomerism plays an important role. A detailed review of these topics is given in Ref. 104. Experimentally, nucleobases are well characterized in the gas-phase. For guanine in the gas phase, the dominance of the 7N-keto tautomer is shown in experiment and theory, while in solution the 9N-keto tautomer, known from Watson-Crick base pairing, is the most abundant.[25,105,106] For adenine, too, the 9N-tautomer is the main form, with minor fractions of the N7- and N3-tautomers, dependent on the environment. In the gas-phase, imino-tautomers are observed as well.[106,107] Cytosine has three different tautomers, the Watson-Crick tautomer, the 3N-tautomer, and the enol tautomer. While in the gas-phase, in argon and nitrogen matrices, multiple tautomers occur, in solution only keto-tautomers are present, with the Watson-Crick tautomer being the most abundant.[105,106] Thymine and uracil exist mainly in the keto-form, but some experiments, using derivatives, indicate the presence of a small amount of enol-tautomers even in solution.[106] For all nucleobases, it is important to also consider the ionization state, since the removal or addition of a proton drastically increases the number of accessible protonation patterns. A lot of experimental and theoretical studies also cover the base hypoxanthine and the respective nucleoside inosine. These substances are not investigated in this work; the tautomers investigated here are described in more detail in Table 20 in chapter *4.5.1*.

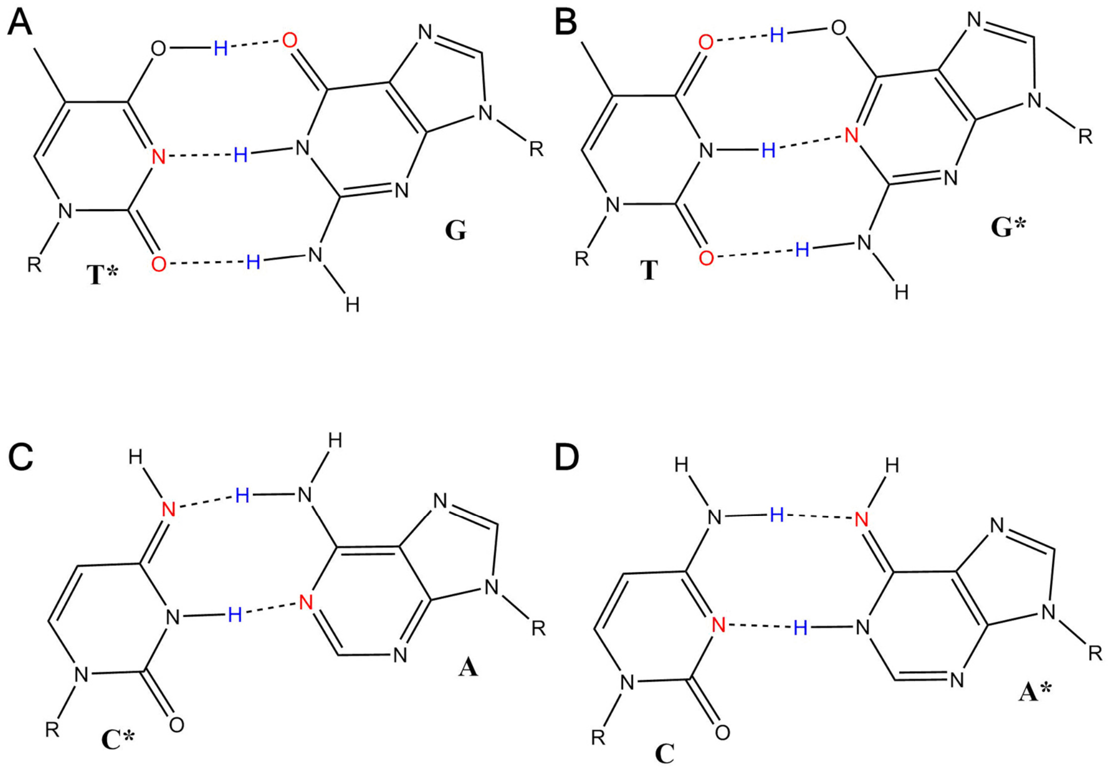


Figure 4: Mispairings of DNA nucleobases involving rare tautomers. Two thymine-guanine base pairs are possible (A and B) involving keto-enol tautomers of the thymine (A) or guanine (B). Two cytosine-adenine pairs are also possible (C and D) involving amino-imino tautomers of the cytosine (C) or adenine (D). Hydrogen bond donors are marked in blue, acceptors in red.

1.7 Non-natural nucleic acids – the Hachimoji code

Natural DNA satisfies the central dogma of molecular biology. An example for an artificial nucleic acid system is the Hachimoji code (from the japanese words for eight “hachi” and letter “moji”). Hachimoji DNA and RNA exhibit catalytic activity, and transcription from DNA to RNA can be performed. By introducing additional letters, the information density is increased compared to the natural nucleic acids.[108] The code consists of the additional purine like bases P (5-aza-7-deazaguanine) and B (isoguanine or 2-hydroxyadenine) as well as the pyrimidine like bases Z (6-amino-5-nitropyridin-2-one) and S, for which two variants exist, one for Hachimoji DNA, the dS (1-methylcytosine), and one for Hachimoji RNA, the rS (isocytosine), similar to the thymine and uracil split between natural DNA and RNA. In contrast to the natural nucleic acids, the Hachimoji base pairs all are held together by three hydrogen bonds; these base pairs are shown in Figure 5.

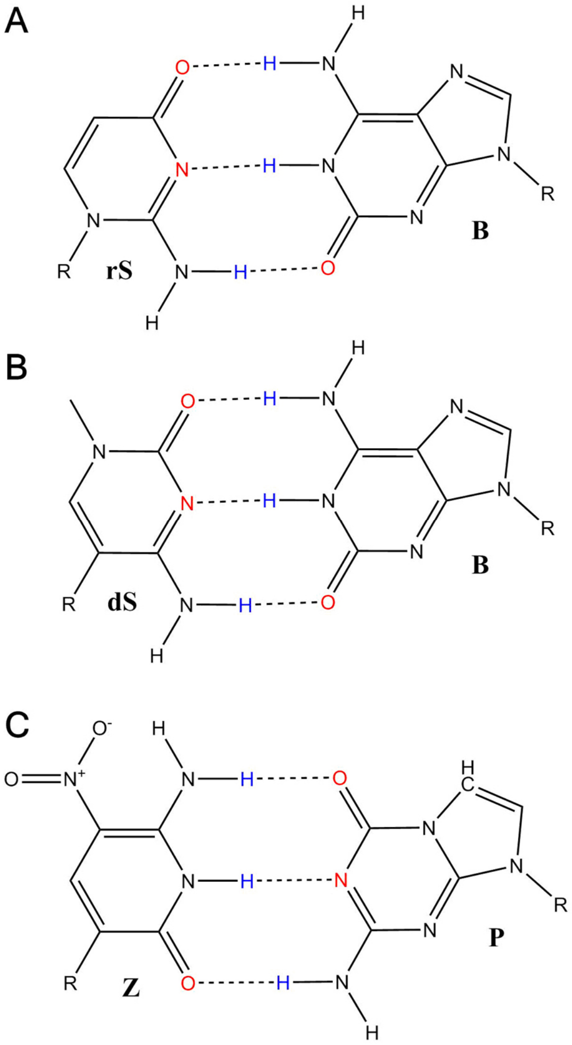


Figure 5: Hachimoji base pairs. The isocytosine-isoguanine pair used for Hachimoji RNA (A) and the Hachimoji DNA pairs 1-methylcytosine-isoguanine (B) and 6-amino-5-nitropyridin-2-one-5-aza-7-deazaguanine (C). Hydrogen bond donors are marked in blue, acceptors in red.

An interesting aspect of the Hachimoji code is the fact that the bases are purine and pyrimidine-like. While

B and S are “real” purine and pyrimidine bases, P and Z are derivatives by leaving a nitrogen atom out of the pyrimidine in case of Z and changing the position of a nitrogen atom from 7 to 5 in P. By using these nucleobase scaffolds and building pairs of a large and a small base, the overall geometry of natural nucleic acids is mimicked, resulting in similar geometries of the double helix and an overall structure close to natural DNA. Even native T7 RNA polymerase is able to transcribe the P-Z base pair in both directions and the dS to B pair and only fails to incorporate the rS opposite of B. Using a mutant, the FAL T7 RNA polymerase, the transcription of the whole Hachimoji code is possible. The Hachimoji code is able to undergo Darwinian evolution.[3] The required mutation events may occur from tautomerization events; in Figure 6, mispairings of Hachimoji P and B tautomers with the natural thymine are shown. The investigation of the tautomer stability of Hachimoji bases is a major part of this work.[9] All of the tautomers investigated in this work are presented in chapter 4.5 Table 20.

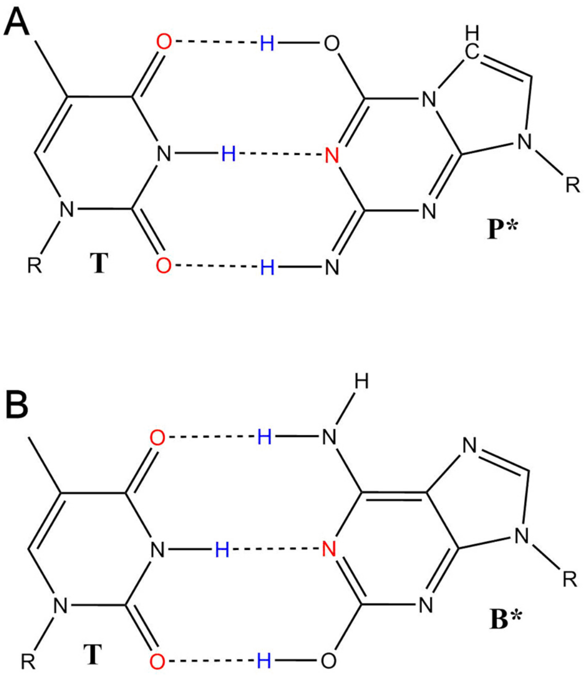


Figure 6: Mispairings of tautomers of the Hachimoji purines P (A) and B (B) with the natural pyrimidine base thymine; similar mispairings are possible with uracil.

1.8 Thesis outline

In the next section of this work, the theory and the methods used are outlined. This includes quantum chemical as well as molecular mechanics-based methods with a focus on solvation modelling, conformational sampling and the calculation and measurement of NMR chemical shifts. Afterwards, the computational and experimental protocols used for the data acquisition are outlined. In the following results section, at first an extension for temperature variation for the EC-RISM solvation model is introduced, this is followed by benchmarking of the predictivity of EC-RISM for tautomer ratios using the SAMPL2 dataset. Subsequently histamine, a model system with multiple ionization states, tautomers and conformations is investigated in detail. Thereby, all new developed methods are applied to benchmark them on a well-known test system. In the context of this benchmarking, a new approach for the optimization of force-field (FF) parameters is introduced and, most importantly, the performance of a combined computational and NMR spectroscopic approach for the prediction of tautomer ratios is explored. This approach is based on the calculation of NMR chemical shifts for the tautomers and a fitting of them onto experimental chemical shifts of the respective ionization states. An important aspect of this chapter is the determination of the nuclei most sensitive for tautomerism; the ^{15}N nucleus is the most important one for the determination of nitrogen heterocycle tautomerism. For computational NMR spectroscopy, shielding constants of reference substances are needed for each nucleus of interest. Such reference shielding constants are introduced for the use in pressure and temperature dependent EC-RISM calculations. All these developments are afterwards used for the investigation of the tautomer stability of nucleic acid building blocks. Thereby, the systems increase in complexity, from nucleobases, natural and non-natural species are considered with a focus on the Hachimoji code, to investigations on the conformational preferences of the AMP nucleotide at high pressures in a two state system, and finally, the examination of the temperature dependent tautomerism of all natural nucleosides and nucleotides by consideration of the conformational ensemble and applying the combined computational and NMR spectroscopic approach. Therefore, the chemical shifts of all nuclei of the base of the nucleotides are measured and calculated temperature dependent.

II. Theory

2.1 Basics of the nuclear magnetic resonance spectroscopy

In chapter 1.3 the importance of NMR spectroscopy for the tautomer elucidation is described. This chapter will explain the basics of this spectroscopic method. The basis of NMR is the nuclear angular momentum *P* which is defined as[32]

$$P = \sqrt{I\left(I+1\right)}\hbar\,, \tag{1}$$

where *I* is the nuclear angular momentum quantum number, often called nuclear spin, and $\hbar$ the reduced Planck constant. This number equals zero for atomic nuclei with an even number of protons and neutrons (these nuclei cannot be observed by NMR), an non-zero integer if the nucleus has an odd number of protons and neutrons, and half-numbered if the number of protons or neutrons is odd.[109] The angular momentum and the magnetic moment $\boldsymbol{\mu}$ of the nucleus are connected via the gyromagnetic ratio γ, which is different for each isotope

$$\boldsymbol{\mu} = \gamma\mathbf{p}\,. \tag{2}$$

During an NMR experiment, the nuclear angular momentum is influenced by an external magnetic field *B*; the strength in the flux direction z of this field is

$$p_z = m\hbar\,, \tag{3}$$

where *m* is the magnetic quantum number, resulting in a splitting into 2*I*+1 energy states, one for each possible orientation[110]

$$E = -m\gamma\hbar B_0\,, \tag{4}$$

with the flux density B_0 of the magnetic field, which is called Zeeman effect. The nuclear spins undergo the so-called Larmor precession with the frequency ν_L, leading to an energetic difference ΔE between the Zeeman levels[32]

$$\nu_L = \left|\frac{\gamma}{2\pi}\right| B_0\,, \tag{5}$$

$$\Delta E = \gamma\hbar B_0\,. \tag{6}$$

These levels are occupied according to Boltzmann statistics, some of the most observed NMR nuclei like

^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P have a magnetic quantum number of $m_I = \pm 1/2$, yielding a ratio of

$$\frac{N_{(m=-1/2)}}{N_{(m=+1/2)}} = \exp(-\beta\Delta E), \tag{7}$$

their spins can be excited using radiation of the energy $h\nu = \Delta E$. β is the inverse temperature: $(k_\mathrm{B}T)^{-1}$ with the Boltzmann constant k_B, respectively the gas constant R for molar quantities. The needed frequency depends on the local magnetic field B_eff at the nucleus which is given by

$$B_\mathrm{eff} = B_0 - \sigma B_0 , \tag{8}$$

with the shielding constant σ. Using this local magnetic field, the resonance frequency is

$$\upsilon_0 = \frac{\gamma}{2\pi} B_0 (1-\sigma). \tag{9}$$

The local magnetic field depends on the surroundings of the nucleus and, thus, allows for conclusions about the chemical environment. The resonance frequency depends on the strength of the external magnetic field which varies for different spectrometers, needing for a standardization of the spectra. Because of that, a reference substance is needed, this substance is in most cases (according to IUPAC) tetramethylsilane (TMS) for ^{1}H and ^{13}C nuclei and liquid ammonia for ^{15}N; in water, the more soluble sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) is used.[111] The resulting standardized values are called chemical shifts and are used for example for structure elucidation.[110]

A nucleus can also interact with a neighboring nucleus; this is called nuclear coupling. Coupling is possible via space or indirectly via the bond electrons. The mechanism of the so-called scalar coupling (*J*-coupling) is based on the different spins of the bond electrons, which are influenced to different degrees by the spin of the respective nuclei, resulting in a splitting of the NMR signal. The number of bonds involved in the scalar couplings is typically denoted, i.e. ^2J denotes the scalar coupling constant via two bonds. The nuclear Overhauser (NOE) effect exploits the so-called dipolar, or direct, coupling through space, which normally cancels in solution, by saturating the resonance of a nucleus. The relaxation rate of the dipolar coupling depends on the distance of the nuclei, which is measured via the signal intensity and allows the calculation of nuclear distances. Both effects will be mentioned in this work, but play a minor role compared to the chemical shifts. For a detailed description of *J*-couplings and NOEs the usual literature, like Ref. 32, is recommended.

2.2 Basics of quantum chemical calculations

In quantum chemistry, each system can be assigned to a wave function that describes it completely. With the help of suitable operators, the desired observables can be extracted from this wave function. In this work, only the time-independent, non-relativistic case is considered, in which the Schrödinger equation, that describes the relationship between the wave function Ψ and the energy E via the molecular Hamilton operator $\hat{H}$ (which can be divided into the kinetic $\hat{T}$ and potential $\hat{V}$ energy operators), reads as follows:[112]

$$\hat{H}\Psi = E\Psi \,. \tag{10}$$

Analytically, the Schrödinger equation can only be solved for simple systems like the hydrogen atom, more complex systems need the introduction of numerical solution techniques and approximations. In this work, the Born-Oppenheimer approximation,[113] in which electrons and nuclei are treated differently due to the large difference in mass, is used. This approximation allows the description of the nuclei using classical mechanics potentials since the motion of the nuclei is only slightly influenced by the electrons and negligible; an instantaneous relaxation of the electrons, having a much higher speed and lower mass, is assumed, allowing for a separation of the wave function into and an electronic and a nuclear part:[112]

$$\Phi(\{\mathbf{r}_i\};\{\mathbf{R}_A\}) = \Psi(\{\mathbf{r}_i\};\{\mathbf{R}_A\})\eta(\{\mathbf{R}_A\}) \,. \tag{11}$$

Now, there is only a parametric dependence of the electronic wave function $\Psi(\{\mathbf{r}_i\};\{\mathbf{R}_A\})$ on a given set of nuclear coordinates $\{\mathbf{R}_A\}$, and that allows a numerical solution with less computational effort, and together with the nuclear wave function $\eta(\{\mathbf{R}_A\})$, the total wave function $\Phi(\{\mathbf{r}_i\};\{\mathbf{R}_A\})$ can be calculated; $\{\mathbf{r}_i\}$ are the electronic coordinates. The following chapters *2.2.1-2.2.6* describe different numerical techniques used for quantum chemical calculations in this work.

2.2.1 The Hartree-Fock method

For all wave function methods, like Hartree-Fock, the namesake function is needed, also in most cases, the Born-Oppenheimer approximation is used. This approximation is in detail explained in chapter 2.2 and *2.2.4*. Although the wave function is unknown at the beginning of the calculation, it can be approximated using the linear combination of atomic orbitals (LCAO) technique. Here, the wave function is constructed as a linear combination of known wave functions, so called basis functions ϕ_m, with coefficients c_m, which can be the analytically accessible wave functions of the hydrogen atom or approximations of them (in more detail discussed in chapter *2.2.6*)

$$|\psi\rangle = \sum_{\mathrm{m}} c_{\mathrm{m}} |\varphi_{\mathrm{m}}\rangle \, . \tag{12}$$

Using the variational principle (Rayleigh-Ritz method), the coefficients can be optimized; the optimal wave function is the one with the lowest ground state energy eigenvalue E_0:[112]

$$\langle\psi|\hat{H}|\psi\rangle \geq E_0 \, . \tag{13}$$

With real basis functions and by applying the Lagrangian formalism, this problem can be rewritten as a matrix equation, with the Hamilton matrix **H**, overlap matrix **S**, and coefficient vector **c**; the resulting equation and the matrix elements are given as[62,114]

$$\mathbf{Hc} = E\mathbf{Sc} \, , \tag{14}$$

$$H_{ij} = \langle\varphi_i|\hat{H}|\varphi_j\rangle \, , \tag{15}$$

$$S_{ij} = \langle\varphi_i|\varphi_j\rangle \, . \tag{16}$$

This is the general matrix equation, called characteristic equation, for wave function approaches. For multi-electron problems, the Pauli exclusion principle and the spin, α or β, of the electrons have to be considered; two electrons can occupy the same position in space when having different spin functions, so there are two spin orbitals χ for every space orbital[115]

$$\chi(\mathbf{x}) = \begin{cases} \phi(\mathbf{r})\alpha(\omega) \\ \phi(\mathbf{r})\beta(\omega) \end{cases} , \tag{17}$$

where the position coordinates **r** and spin coordinates ω are independent. These spin orbitals can be combined to a determinant, which is a universal multi-electron wave function, called Slater determinant, following the Pauli principle[115]

$$\Psi(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_i(\mathbf{x}_1) & \chi_j(\mathbf{x}_1) & \cdots & \chi_k(\mathbf{x}_1) \\ \chi_i(\mathbf{x}_2) & \chi_j(\mathbf{x}_2) & \cdots & \chi_k(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(\mathbf{x}_N) & \chi_j(\mathbf{x}_N) & \cdots & \chi_k(\mathbf{x}_N) \end{vmatrix} . \tag{18}$$

For a given number of electrons, all possible Slater determinants can be constructed. In practice, not all of them have to be calculated due to symmetry reasons. One important approximation in the Hartree-Fock method is the use of only a single determinant, the Hartree-Fock ground state (all excited states are neglected), the other the reduction of a many-body problem to an effective single-body problem. This is done using a one-electron operator, the Fock operator $\hat{F}$, for electron *i* while the effect of all other electrons *j* is approximated by an average effect (here for a restricted calculation and in atomic units)

$$\hat{F}(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{\mathrm{j}=1}^{N/2}\left[2\hat{J}_{\mathrm{j}}(i) - \hat{K}_{\mathrm{j}}(i)\right]. \tag{19}$$

Here the first two terms are the single-electron Hamilton operator, ∇ the Nabla operator, Z_A the proton number of nucleus *A*, r_{iA} the distance between *A* and electron *i*, *N* is the number of electrons (in a closed-shell system the number of occupied orbitals is half the number of electrons), $\hat{J}$ the Coulomb operator and $\hat{K}$ the exchange operator.[115] Because of this approximations, HF is often called mean-field approximation. The Schrödinger equation, using the Fock- instead of the Hamilton operator, can now be solved iteratively, mostly using fixed-point iterations, until convergence; this is called a self-consistent field method (SCF).[62]

2.2.2 Møller-Plesset perturbation theory

The many-body wave function that can be constructed from the Slater determinants (Eqn. 18) using a linear combination of all possible determinants is called full configuration interaction (FCI) wave function. The difference between the FCI and HF energies is called correlation energy. FCI calculations are computationally demanding and time consuming and often impossible; therefore, it is important to approximate the correlation energy. An important method to do this is the Møller-Plesset perturbation theory (MP*n*, where *n* is the order of perturbation, being 2 in this work),[63] which is based on the Rayleigh-Schrödinger perturbation theory. Here, the unperturbed Hamilton operator is extended by a perturbation term $\lambda\hat{V}$ resulting in a power series expression of the wave function and energy[114]

$$\left|\Psi_n\right\rangle = \lim_{n\to\infty}\sum_{i=0}^{n}\lambda^i\left|\Psi_n^{(i)}\right\rangle, \tag{20}$$

$$E_n = \lim_{n\to\infty}\sum_{i=0}^{n}\lambda^i E_n^{(i)}. \tag{21}$$

Applying the perturbed Hamilton operator to the unperturbed wave function from a HF calculation results in the following expression for the energy (the order of perturbation equals the order of the power series)

$$(\hat{H} + \lambda\hat{V})(\left|\Psi_n^{(0)}\right\rangle + \lambda\left|\Psi_n^{(1)}\right\rangle + \lambda^2\left|\Psi_n^{(2)}\right\rangle + ...) \\ = (E_n^{(0)} + \lambda E_n^{(1)} + \lambda^2 E_n^{(2)} + ...)(\left|\Psi_n^{(0)}\right\rangle + \lambda\left|\Psi_n^{(1)}\right\rangle + \lambda^2\left|\Psi_n^{(2)}\right\rangle + ...) . \tag{22}$$

Here, the first order perturbation energy is the HF energy by definition, while the second-order perturbation (MP2) energy is given by[114]

$$E_0^{(2)} = \sum_{\substack{a<b \\ r<s}} \frac{\left|\langle ab\|rs\rangle\right|^2}{\varepsilon_a + \varepsilon_b - \varepsilon_r - \varepsilon_s}, \tag{23}$$

with the eigenvalues ε_a, ε_b, ε_r and ε_s for states $|a\rangle$, $|b\rangle$, $|r\rangle$ and $|s\rangle$; $|ab\rangle$ is the ground state, and $|rs\rangle$ the second excited state. MP*n* theory can easily be applied to restricted HF calculations; open-shell systems may cause some problems but are not investigated in this work.

2.2.3 Coupled Cluster Theory

The coupled-cluster (CC) theory is an alternative way to approximate the correlation energy, which is based on the many-body perturbation theory.[65] Like for MP*n*, a HF calculation has to be done prior to the CC calculation (in theory both would work also with other reference wave functions like from CI calculations, but HF is the standard) to get access to the ground state wave function Ψ_0, which is expanded to the CC wave function Ψ with the wave operator $e^{\hat{t}}$ [66,115]

$$\Psi = e^{\hat{t}}\Psi_0 . \tag{24}$$

The wave operator is given by a Taylor-series expansion

$$e^{\hat{t}} \equiv 1 + \hat{t} + \hat{t}^2/2! + \hat{t}^3/3! + ... = \sum_{\mathrm{k}=0}^{\infty} \hat{t}^k/k!, \tag{25}$$

where the cluster operator $\hat{t}$ is given by a series of connected operators

$$\hat{t} = \hat{t}_1 + \hat{t}_2 + ... + \hat{t}_N . \tag{26}$$

The components of the cluster operator consist of creation $\hat{c}^a$ and annihilation $\hat{c}_i$ operators,

$$\hat{t}_N = 1/(N!)^2 \sum_{i_1,i_2,\cdots,i_N} \sum_{a_1,a_2,\cdots,a_N} t_{i_1,i_2,\cdots,i_N}^{a_1,a_2,\cdots,a_N} \hat{c}^{a_1}\hat{c}^{a_1}\cdots\hat{c}^{a_N}\hat{c}_{i_N}\cdots\hat{c}_{i_2}\hat{c}_{i_1} \tag{27}$$

where *i* and *a* are the occupied and unoccupied orbitals, respectively. t_i^a denote numerical coefficients, determination of them is needed for construction of the CC wave function, which has to be normalized at the end of the calculation.[115] The creation and annihilation operators will create all excited states needed up to the excitation order determined by the cluster operator. Like MP*n*, CC is non-variational, but size consistent. In this work, the coupled-cluster calculations are done using single and double excitations while neglecting higher orders in Eqn. 26. Besides, a correction for triple excitations using perturbation theory as described in chapter *2.2.2* is applied; this is called CCSD(T). A problem of CC calculations is the high computational cost that scales with a power of seven with respect to the number of basis functions, and a comparatively large basis set is required. There are several approximations which improve the scaling behavior of correlated methods like MP*n* or CC. In this work, the RI/F12 approximations are used. The resolution of the identity (RI, often also called density fitting in the context of DFT) aims to accelerate the expensive calculation of the Coulomb interactions, especially the electron-electron repulsion integrals (ERI) and is available in a wide range of implementations and algorithms, from the early developments (Ref. 116 and 117) to modern algorithms even for periodic systems (Ref. 118, 119 and 120). The original Coulomb matrix from an LCAO ansatz is given by

$$J_{\mu,\nu} = \sum_{\kappa,\tau} P_{\kappa,\tau} \left\langle \mu\nu \middle| \kappa\tau \right\rangle = \sum_{\kappa,\tau} P_{\kappa,\tau} \iint \frac{\phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\phi_\kappa(\mathbf{r}_2)\phi_\tau(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 , \tag{28}$$

with the one-particle density matrix **P** and the ERI as the double integral, having four indices and two centers which makes the evaluation computational demanding and costly. RI introduces auxiliary fitting basis functions η which can be used to approximate the electron density via a linear combination (density fitting) or the Coulomb matrix[119,120]

$$J_{\mu,\nu} = \sum_r d_r \left\langle \mu\nu \middle| r \right\rangle , \tag{29}$$

where the d_r are fit coefficients obtained by

$$\mathbf{Vd} = \mathbf{g} . \tag{30}$$

Here, **V** is a positive definite matrix of the ERIs over the auxiliary basis functions, and **g** the contraction vector of the density matrix composed of the three index ERIs

$$V_{r,s} = \left\langle r \middle| s \right\rangle = \iint \frac{\eta_r(\mathbf{r}_1)\eta_s(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 , \tag{31}$$

$$g_r = \sum_{\mu,\nu} P_{\mu,\nu} \left\langle \mu\nu \middle| r \right\rangle = \sum_{\mu,\nu} P_{\mu,\nu} \iint \frac{\phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\eta_r(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 . \tag{32}$$

The dimensionality of the integrals is now reduced and Eqn. 30 can be solved efficiently using the Choleksy decomposition and optimized computational algorithms.[119,121] For the RI approximation, the auxiliary basis sets, corresponding to the respective basis set used for the calculation, have to be used; the generation of these basis sets is in detail discussed in Ref 122,123 and 124.

The F12 correction is based on its precursor, the R12 approximation[125]. In these methods, explicit correlation effects are only calculated for doubly occupied orbitals, therefore the wave function is expanded using geminals, which are pair functions of the following form

$$\hat{Q}_{12} r_{12} \varphi_i(1) \varphi_j(2), \tag{33}$$

$$\hat{Q}_{12} e^{\gamma r_{12}} \varphi_i(1) \varphi_j(2). \tag{34}$$

The first term is used in R12 theory,[125] the second in F12 theory;[126] the projector operator $\hat{Q}$ is given by

$$\hat{Q}_{12} = \left(1 - \hat{O}_1\right)\left(1 - \hat{O}_2\right)\left(1 - \hat{V}_1 \hat{V}_2\right), \tag{35}$$

$$\hat{O}_1 = \sum_i \left|\varphi_i(1)\right\rangle \left\langle \varphi_i(1)\right|, \tag{36}$$

$$\hat{V}_1 = \sum_a \left|\varphi_a(1)\right\rangle \left\langle \varphi_a(1)\right|, \tag{37}$$

where the indices *i* and *j* indicate occupied and *a*, *b* virtual orbitals.[127] Using this operator, the wave function can be rewritten with pair functions *u* as matrix elements with the basis functions $\phi_{\alpha,\beta}$ (here for R12, the complementary equation for F12 can be built from Eqn. 34)

$$u_{kl}^{\alpha\beta} = \left\langle \varphi_\alpha \varphi_\beta \middle| \hat{Q}_{12} r_{12} \middle| \varphi_k \varphi_l \right\rangle. \tag{38}$$

To avoid the calculation of higher order integrals, the R12/F12 theories are most often used with the RI approximation, and therefore auxiliary basis functions are needed to approximate these matrix elements. To calculate the R12/F12 approximations, nowadays so-called complementary auxiliary basis sets (CABS) are used. More details of the generation and implementation of these basis can be found in Ref 128 and 129. In this work, CCSD(T) calculations using the RI and F12 approximations are performed using the Orca program package.[130]

2.2.4 Density functional theory in quantum chemistry

The quantum chemical density functional theory (DFT) is based on the Hohenberg-Kohn theorem.[67] It is proven that the ground-state electron density uniquely determines ground-state properties such as the ground-state wave function, energy and thus all other electronic properties of the molecule. The electronic Hamilton operator using the Born-Oppenheimer approximation is[115]

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{j}^{N}\sum_{i>j}^{N}\frac{1}{r_{ij}}, \tag{39}$$

where the coordinates of the nuclei are fixed, so that the second term, describing the electron-nuclei interactions, can be rewritten as a summation over all electrons, influenced by the external potential of the nuclei, now depending only on the electron coordinates

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N}v(\mathbf{r}_i) + \sum_{j}^{N}\sum_{i>j}^{N}\frac{1}{r_{ij}}, \tag{40}$$

$$v(\mathbf{r}_i) = -\sum_{A=1}^{M}\frac{Z_A}{r_{iA}}. \tag{41}$$

According to the Hohenberg-Kohn theorem, this external potential as well as the number of electrons are determined by the ground-state electron probability density ρ_0, so that (knowing the number of electrons and the external potential, the dependence of the ground state energy E_0 from the external potential is stated with the formalism $E_v[\rho_0]$) the ground-state wave function ψ_0 can be expressed as a functional of this density[115]

$$E_0 = E_v\left[\rho_0(\mathbf{r})\right] = E_0\left[\sum_{i=1}^{N}\left|\psi_{0,i}(r)\right|^2\right]. \tag{42}$$

The electronic Hamiltonian, as the sum of three terms, electronic kinetic energies and electron-nuclei (Ne)/electron-electron (ee) potential energies, can be written as sum of the average values determined by the ground-state wave function and consequently ρ_0, as a functional

$$E_0 = E_v\left[\rho_0\right] = \bar{T}\left[\rho_0\right] + \bar{V}_{Ne}\left[\rho_0\right] + \bar{V}_{ee}\left[\rho_0\right] = \int \rho_0(\mathbf{r})v(\mathbf{r})d\mathbf{r} + \bar{T}\left[\rho_0\right] + \bar{V}_{ee}\left[\rho_0\right]. \tag{43}$$

The problem here is that the functionals $\bar{T}\left[\rho_0\right]$ and $\bar{V}_{ee}\left[\rho_0\right]$ are unknown, leading to the Hohenberg-Kohn variational theorem[115]

$$E_v\left[\rho_0\right] \le \int \rho_{\mathrm{tr}}(\mathbf{r})v(\mathbf{r})d\mathbf{r} + \bar{T}\left[\rho_{\mathrm{tr}}\right] + \bar{V}_{ee}\left[\rho_{\mathrm{tr}}\right], \tag{44}$$

where ρ_{tr} is a trial density. The step to overcome the problem of unknown functionals is the Kohn-Sham method, here the unknown functionals are rewritten using functionals for a reference system of non-interacting particles, denoted with the subscript s[68]

$$\Delta\bar{T}[\rho_0] \equiv \bar{T}[\rho_0] - \bar{T}_s[\rho_0], \tag{45}$$

$$\Delta\bar{V}_{ee}[\rho_0] \equiv \bar{V}_{ee}[\rho_0] - \frac{1}{2}\iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 \,. \tag{46}$$

The latter term describes the electrostatic repulsion energy for electrons within a continuous electron charge distribution within the density ρ. Inserting both of these equations in Eqn. 44 yields

$$E_0 = E_v[\rho_0] = \int \rho_0(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} + \bar{T}_s[\rho_0] + \frac{1}{2}\iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 + E_{xc}[\rho_0] \tag{47}$$

with the exchange-correlation energy functional $E_{xc}[\rho_0]$, which is the last unknown functional and has to be approximated

$$E_{xc}[\rho_0] \equiv \Delta\bar{T}[\rho_0] + \Delta\bar{V}_{ee}[\rho_0]. \tag{48}$$

In this work, a functional from the group of Becke's three-parameter hybrid functionals is used, Becke-3-Lee-Yang-Paar (B3LYP),[131,132] where the exchange-correlation functional is given by:[133]

$$E_{XC}^{\mathrm{B3LYP}} = E_x^{\mathrm{LDA}} + a_0(E_x^{\mathrm{HF}} - E_x^{\mathrm{LDA}}) + a_x(E_x^{\mathrm{GGA}} - E_x^{\mathrm{LDA}}) + E_c^{\mathrm{LDA}} + a_c(E_c^{\mathrm{GGA}} - E_c^{\mathrm{LDA}})\,. \tag{49}$$

Hybrid functionals are using the exchange energy from HF calculations, which is denoted as E_x^{HF}; a_0, a_{x} and a_{c} are empirical parameters, E_x^{LDA} and E_c^{LDA} are the exchange and correlation energies from the Vosko-Wilk-Nusair functional[134] using the local-density-approximation (LDA), E_x^{GGA} the exchange energy from the Becke-88 exchange functional[131] and E_c^{GGA} the correlation energy from the Lee-Yang-Parr correlation functional,[132] the latter using the generalized-gradient-approximation (GGA).

2.2.5 Calculation of nuclear magnetic resonance parameters

In chapter 1.3 and 1.4, the importance of NMR parameters from experiment and computation is pointed out. In this work, the focus lies on the shielding constants. The theory of NMR parameter calculation is mainly developed by Ramsey in a series of papers.[135,136] The parameters can be calculated using the second order Rayleigh-Schrödinger perturbation theory (for an electronic system modified by a perturbation **x**) for stationary, non-degenerated systems[137]

$$\frac{\partial^2 E(\mathbf{x})}{\partial x_i \partial x_j} = \left\langle \Psi_0 \middle| \frac{\partial^2 \hat{H}}{\partial x_i \partial x_j} \middle| \Psi_0 \right\rangle - 2\sum_{n\neq 0} \frac{\left\langle \Psi_0 \middle| \partial \hat{H} / \partial x_i \middle| \Psi_n \right\rangle \left\langle \Psi_n \middle| \partial \hat{H} / \partial x_j \middle| \Psi_0 \right\rangle}{E_n - E_0} \tag{50}$$

This equation describes the mixed second order derivatives of the energy and can be used for all observables which depend on the energy with the second order (for example electron paramagnetic resonance (EPR) parameters). It consists of the average of the ground-state derivative and the sum over the excited states of Ψ_n with a normalization based on the energy E_n of the respective state. The chemical shift of a system is, as described in chapter 2.1, dependent on the shielding of a reference substance. Using the tensor formalism, it is given by[138]

$$\boldsymbol{\delta} = \mathbf{1}\sigma_{\text{ref}} - \boldsymbol{\sigma} . \tag{51}$$

Here **1** is the identity matrix and δ the chemical shift tensor, for a shielding tensor **σ** using a reference shielding $\sigma_{\text{ref,}}$; the trace of the shielding tensor is the shielding of the system[138]

$$\sigma = \frac{1}{3}\text{Tr}(\boldsymbol{\sigma}) . \tag{52}$$

In this work, only the isotropic chemical shift is investigated, the anisotropy is neglected (due to the fast movement and rotation of a molecule in solution, these contributions average out; a detailed discussion about this phenomenon is given in the later chapters).[139] It is given by[135,137]

$$\boldsymbol{\sigma} = \frac{\partial^2 E}{\partial \boldsymbol{\mu}_A \partial \mathbf{B}} = \left\langle \Psi_0 \middle| \hat{\mathbf{H}}^{\text{DS}} \middle| \Psi_0 \right\rangle - 2\sum_{n\neq 0} \frac{\left\langle \Psi_0 \middle| \hat{\mathbf{H}}^{\text{PSO}} \middle| \Psi_n \right\rangle \left\langle \Psi_n \middle| \hat{\mathbf{L}}_{\text{G}} \middle| \Psi_0 \right\rangle}{E_n - E_0} . \tag{53}$$

The shielding is the mixed second derivative of the magnetic momentum of the nucleus $\boldsymbol{\mu}_A$ and the external magnetic induction **B**. For the calculation, three operators are needed, the diamagnetic shielding (DS) operator $\hat{\mathbf{H}}^{\text{DS}}$, the paramagnetic spin orbit (PSO) operator $\hat{\mathbf{H}}^{\text{PSO}}$, and the orbital Zeeman operator $\hat{\mathbf{L}}_{\text{G}}$ [140]

$$\hat{\mathbf{L}}_{\text{G}} = \frac{1}{2}\mathbf{r}_{iG} \times \hat{\mathbf{l}}_i , \tag{54}$$

$$\hat{\mathbf{H}}_A^{\text{PSO}} = \alpha^2 \sum_i \frac{\hat{\mathbf{l}}_{iA}}{r_{iA}^3} , \tag{55}$$

$$\hat{\mathbf{H}}_{\mathbf{B}A}^{\text{DS}} = \frac{\alpha^2}{2} \sum_i \frac{\mathbf{r}_{iG}\mathbf{r}_{iA} - \mathbf{r}_{iA}\mathbf{r}_{iG}^T}{r_{iA}^5} , \tag{56}$$

where the superscript T denotes triplet states. There are similar expressions for the other NMR parameters;

for example the spin-spin coupling constants can be calculated as the mixed second derivative of the energy with respect to the magnetic moments of both interacting nuclei.[141,142] These operators consist of the fine-structure constant α, the angular momentum operator $\hat{\mathbf{l}}$ and the electron coordinates with respect to the nucleus A, $\mathbf{r}_{iA}$ and, more difficult, the origin of the gauge G, $\mathbf{r}_{iG}$. This gauge origin is influencing the paramagnetic contributions via $\hat{\mathbf{L}}_{\mathrm{G}}$ and the diamagnetic contributions due to $\hat{\mathbf{H}}^{\mathrm{DS}}$.[140] This dependence cancels for the complete basis set and decreases with larger basis sets,[143] but for real calculations another solution of the problem is needed due to the high computational cost of large basis sets (especially using MP2, because the frozen core approximation cannot be used for NMR calculations). There are multiple methods to overcome this issue, like individual gauge for localized orbitals (IGLO),[47] localized orbitals local origin (LORG),[144,145] individual gauges for atoms in molecules (IGAIM),[146] and continuous gauge transformation (CSGT),[147,148] but the most used one, also used for this work, is the method of gauge invariant atomic orbitals (GIAO, also called gauge-including or gauge-dependent atomic orbital or London orbital)[149,150,151,152]

$$\chi_i = \exp(-\left(i/c\right)\mathbf{A}_i \cdot \mathbf{r})\varphi_i \,. \tag{57}$$

The GIAO χ_i is composed of the basis function ϕ_i (with the corresponding coefficient c) and a vector potential $\mathbf{A}_i$ (the first i in the exponential is the imaginary unit, not the electron index), making the wave function dependent on the gauge origin but the magnetic property independent of it, allowing for the calculation of NMR parameters gauge invariant and using incomplete basis sets,[149] for example using coupled perturbed Hartree-Fock theory.[143] A general vector potential has a rotation and a divergence, both are location-dependent. A potential describing a magnetic field is given by[112]

$$\mathbf{B} = \nabla \times \mathbf{A} \,. \tag{58}$$

This potential is differentiated within the calculation; thus it is possible to add a constant which vanishes during differentiation. This constant is the gauge of the potential and can be chosen in a way that the divergence becomes zero; the resulting gauge is called gauge of Coulomb[139]

$$\nabla \cdot \mathbf{A} = 0 \,. \tag{59}$$

Using the nonrelativistic Born-Oppenheimer approximation, the vector potential describing the external and nuclear magnetic field for an electron is given by[138]

$$\mathbf{A}_i\left(\mathbf{r}\right) = \frac{1}{2}\mathbf{B} \times \mathbf{r}_i + \left(\boldsymbol{\mu}_A \times \mathbf{r}_{Ai}\right)/r_{Ai}^3 \,. \tag{60}$$

The full electronic Hamiltonian is given by a Taylor-series expansion[153]

$$\hat{H} = \hat{H}_0 + \sum_{i=1}^{N}\left(\hat{\mathbf{L}}_{\mathrm{G}} \Big|_{\mathbf{B}=0} \mathbf{B} + \sum_{A=1}^{M} \hat{\mathbf{H}}^{\mathrm{PSO}}\Big|_{\boldsymbol{\mu}_A=0} \boldsymbol{\mu}_A + \sum_{A=1}^{M}\hat{\mathbf{H}}^{\mathrm{DS}}\Big|_{\mathbf{B},\boldsymbol{\mu}_A=0}\mathbf{B}\boldsymbol{\mu}_A + \dots\right). \tag{61}$$

Only terms up to the second order have to be considered for the calculation of NMR parameters, since these parameters are second order derivatives of the energy. The calculation of NMR parameters is possible using a variety of quantum chemical programs[151,154,155] and theoretical methods like DFT,[156,157] CC[151,153] or MP2;[153,158] in this work, MP2 is used with the Gaussian program package.[154]

2.2.6 Basis sets

Only the wave function of the hydrogen atom is analytically known, all other wave functions have to be constructed from basis functions using the variational theorem, (Eqn. 12). A set of system-independent, parametrized one-electron basis functions is called basis set. In a recent review article, Nagy and Jensen defined six requirements for a such a basis set:[159]

I. The basis functions should be designed in a way that the orbitals can be represented by a small number of functions

II. The complete basis and therefore a basis set limit should be producible from the basis functions

III. The basis sets should be available at different qualities, allowing for a systematic approach to the basis set limit

IV. Using the basis sets, the calculation of a broad range of molecular properties should be possible

V. The calculation of one- and two-electron integrals should be computationally efficient

VI. The basis set should be available for a large number of atoms

The basis set limit is the limit a computational method can achieve using an infinite (or complete) basis set (CBS). This value is most often extrapolated from a series of calculations using different qualities within the same basis set according to III. Functions that describe the orbitals of the hydrogen atom are the Slater functions, the resulting orbitals are called Slater type orbitals (STO)[160]

$$\varphi_{\zeta,n,l,m}(r,\phi,\theta) = Nr^{n-1}Y_{l,m}\left(\phi,\theta\right)e^{-\zeta r}, \tag{62}$$

with the spherical coordinates r, θ and φ, a normalization constant , Y describes the angular momentum of the STO via a spherical function, ζ is the orbital exponent and n the principal quantum number. While representing the hydrogen orbitals with high accuracy, the STOs have the drawback of high computational effort and therefore cost. This problem can be solved using Gaussian type orbitals (GTO) which can be written in spherical and Cartesian coordinates[114,161]

$$\varphi_{\zeta,n,l,m}(r,\phi,\theta) = Nr^{2n-2-l}Y_{l,m}\left(\phi,\theta\right)e^{-\zeta r^2} , \tag{63}$$

$$\varphi_{\zeta,i,j,k}(x,y,z) = Nx^i y^j z^k e^{-\zeta r^2} . \tag{64}$$

In the Cartesian expression, *x*, *y* and *z* are the coordinates, their powers sum up to the angular momentum of the orbital. Most quantum chemical codes use the Cartesian expression, because integral evaluation is more efficient here.[159] The problem of GTOs is the, compared to STOs and the analytical solution, worse description of the hydrogen orbitals. Using a linear combination of multiple GTOs can overcome this issue at low computational cost because a multiplication of two GTOs can be done by simply summing up the exponents. This results in the STO-*x*G basis sets, where *x* GTOs are combined to approximate an STO. In most basis sets, the functions for the core orbitals are fixed linear combinations because they only differ slightly between isolated atoms and a molecular environment. There are three types of variables for such a basis set, the number of functions of each type of orbital (s-, p-. d-, f-, etc.), the exponents of the functions, and the pre-factors of fixed linear combinations built from them, resulting in a broad variety of available basis sets. The GTOs used for a fixed linear combination are called primitives and the combined functions used in the calculation contractions, respectively. The number of contractions used for the calculation of a valence orbital is called zeta (referring to the exponent of the slater function) and is used for the description of the quality of the basis set.[114,161] Using a basis set with a higher zeta cannot always improve the results of a calculation because most basis sets are optimized for energy calculations. Therefore, basis sets can be augmented to improve the performance in all kinds of calculations. The part of the wave function which is far from the nucleus contributes only slightly to the energy and is often not optimally represented due to the shape of a GTO but is important for example for excited states, anions or the description of hydrogen bonds. Here, augmenting the basis set with so-called diffuse functions is helpful.[159] Adding polarization functions with a higher angular momentum quantum number are helpful for the description of chemical bonds and tight functions can improve the description of properties dependent on the core orbitals, like core electron correlation or magnetic properties such as NMR parameters. The drawback of these augmentations is a higher computational cost.[159,162] Some of the most popular basis sets are the Pople basis sets,[163] made for wave function approaches like HF and with relatively low computational cost, the correlation-consistent or Dunning basis sets,[164,165] including polarization functions for correlated methods, the core-valence basis sets,[166] with a better description of the core electrons, the polarization consistent or Jensen basis sets,[167,168,169] and the Ahlrichs basis sets,[170,171] both of the latter optimized for DFT calculations. It is important to consider that all these basis sets are parametrized without the use of a solvation model. In this work, the Pople basis set 6-311+G(d,p), a triple zeta basis set including polarization functions, is used for B3LYP and MP2 calculations (the frozen core approximation cannot be used for MP2/NMR calculations), and the cc-pVTZ basis set, together with the corresponding auxiliary and complementary auxiliary (CABS) basis sets, is used for CCSD(T)-RI/F12 calculations.

2.3 Classical mechanics

Using the quantum mechanical methods described in chapter 2.2, the atoms are described by their components: electrons and nuclei. These methods yield energetics and NMR parameters with a high level of accuracy. The drawback of these methods is the high computational cost, financially and temporally. Also, these methods are not suited for many applications because the systems are too large for today's computers. The investigation of large protein systems, especially membrane proteins, is not possible using high-level quantum chemistry. Time-resolved trajectories using *ab initio* molecular dynamics simulations are also not available with such systems. To investigate large systems or to sample the conformational degrees of freedom of flexible molecules, methods based on classical mechanics are helpful. These methods will be discussed in the following chapters.

2.3.1 Molecular mechanics

In molecular mechanics, the atoms are not described as electrons and nuclei, they are treated as spherosymmetric particles. These particles are connected by bonds and cannot undergo chemical reactions (there are approaches like polarizable FFs[172] or reactive FFs,[173] but they are not used or discussed in this work). Their interactions with other particles are described by potential functions, composed of terms for bonded atoms as well as long- and short-range interactions for non-bonded atoms. The bonded interactions are harmonic potentials for bonds, bond angles and trigonometric functions for dihedrals, the short range interactions are characterized via Lennard-Jones potentials, the long range interactions using the Coulomb potential.[174,175] There are several of these potential functions, called, together with the respective parameter sets, FFs, with different strengths and weaknesses. Examples are the universal force field (UFF),[176] chemistry at Harvard molecular mechanics (CHARMM),[177,178] ff14SB,[179,180] GROMOS,[181] OPLS,[182] MMFF94s[183] and the general AMBER force field (GAFF)[184] (the latter two used in this work) which is given by

$$U = \sum_{\text{bonds}} k_r (r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{v_n}{2}(1 + \cos(n\varphi - \gamma)) \\ + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right], \tag{65}$$

with the force constants k_r, k_θ, v_n, for bonds with the length r, bond angles with angle θ and dihedrals with multiplicity n, angle ϕ and phase angle γ, respectively; the subscript eq denotes the equilibrium values; the $q_{i,j}$ are the partial charges of atoms i and j at distance R_{ij}, and A and B their Lennard-Jones parameters (they can easily be converted to the more familiar σ and ε).[184] The molecular structure, dependent on coordinates **r**, can be minimized using this potential via the steepest descent or gradient descent method

$$\mathbf{r}_{n+1}(\alpha) = \mathbf{r}_n - \alpha \nabla U(\mathbf{r}_n), \tag{66}$$

with the step size α. The step size can change for each iteration step n, it can be calculated for example as[185]

$$\alpha_n = \frac{\left| (\mathbf{r}_n - \mathbf{r}_{n-1})^T \left[\nabla U(\mathbf{r}_n) - \nabla U(\mathbf{r}_{n-1}) \right] \right|}{\left\| \nabla U(\mathbf{r}_n) - \nabla U(\mathbf{r}_{n-1}) \right\|^2}, \tag{67}$$

this method is used in this work to minimize initial structures of molecules, using the MMFF94s force field and the Avogadro program package.[183,186]

2.3.2 Molecular dynamics simulations

Molecular dynamics (MD) simulations are used to sample the ensemble configuration over time. The FF is used to generate time dependent trajectories, therefore the equations of motions according to Newtonian mechanics have to be solved[174]

$$\mathbf{F}_i = -\nabla U(\mathbf{r}_i) = m_i \mathbf{a}_i, \tag{68}$$

with the force $\mathbf{F}_i$, acting on atom i; m is the mass, and **a** the acceleration, which is the first temporal derivative of the velocity and the second of the atomic coordinates. Therefore, the Newtonian equation of motion is a second order differential equation, which can be approximated using integrators like the Verlet algorithm which is given by a first order Taylor series[174,187]

$$\mathbf{r}_i(t+\Delta t) = -\mathbf{r}_i(t-\Delta t) + 2\mathbf{r}_i(t) + \mathbf{a}_i(t)\Delta t^2 , \tag{69}$$

where t is the current time and Δt the time step, which has to be chosen appropriately; 2 fs is often used. These equations must be solved for every atom in the system, resulting in a large computational effort, especially due to the calculation of all pairwise interactions. This effort can be reduced using a cutoff for the calculation of the short-ranged Lennard-Jones interactions, and using numerical techniques like the Ewald summation for the long-ranged Coulomb interactions. Using Ewald summation, the Coulomb interactions are split into the sum of a short-range term which can be calculated efficiently in real space and a long range term, which is treated in Fourier space (particle mesh Ewald, PME).[188,189] The number of pairwise interactions can also be reduced drastically using implicit solvation models. Another problem is the finite system size in simulations, which would result in strong boundary effects. Using periodic boundary conditions or vacuum boundary conditions helps to overcome this issue. Such MD simulations result in the microcanonical ensemble (*NVE*); using thermostats results in the canonical ensemble (*NVT*), and thermostats and barostats in the isothermal-isobaric ensemble (*NpT*).[174,175] In this work, MD simulations are performed using the program sander from the AMBER 18 software package[190] and the ALPB solvation model (see chapter 2.4.2).

2.4 Solvation models

In chapter 2.2, the complexity of quantum chemical calculations and the high computational cost associated with it are pointed out; CC calculations for example scale in effort with the seventh power to the number of basis functions. If solvent molecules have to be included to calculate the behavior of the molecule in solution, not only many solvent molecules are needed to avoid boundary artifacts, also enough solvent configurations have to be considered to get a sufficient statistical average, like it is done in explicit solvation MD simulations, further increasing the number of basis functions and calculations needed. This is not feasible using contemporary computational resources. Consequently, solvation models have to be used, which are capable of approximating the effect of a bulk solvation on the solute.

2.4.1 Implicit solvation models / Thermodynamic integration

Using implicit solvation models is a way to incorporate the solvation effects on a system without having to incorporate a single solvent molecule. Without solvent molecules, the interactions of the bulk solvent ensemble have to be described implicitly, therefore statistical mechanics approaches are required.[191] In this chapter general concepts of implicit solvation models are introduced, in the following chapters specific solvation models for the use in QC and MM calculations are described. The probability of a specific configuration in a fluctuating solute-solvent system at a given temperature is[192]

$$P\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)=\frac{\exp\left(-\beta U\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)\right)}{\int\exp\left(-\beta U\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}\,\mathrm{d}\,\mathbf{R}_{\mathrm{u}}}, \tag{70}$$

with the probability function P, the potential U, the inverse temperature $\beta=(k_BT)^{-1}$ and the solvent and solute coordinates $\mathbf{R}_{\mathrm{v}}$ and $\mathbf{R}_{\mathrm{u}}$. This potential can be decomposed into the intramolecular solute potential U_{u}, the solvent-solvent potential U_{vv}, and the solute-solvent potential U_{uv}[192]

$$U\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)=U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right), \tag{71}$$

so that the expectation value of a quantity Q dependent only on the solute configuration is given by the weighted average over the probability function[193]

$$\langle Q\rangle=\int Q\left(\mathbf{R}_{\mathrm{u}}\right)P\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}\,\mathrm{d}\,\mathbf{R}_{\mathrm{u}}, \tag{72}$$

resulting in a reduced probability distribution which now only depends on the solute coordinates

$$\bar{P}\left(\mathbf{R}_{\mathrm{u}}\right)=\int P\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}\,. \tag{73}$$

Besides, the expectation value of Q can now be expressed only in solute coordinates

$$\langle Q\rangle=\int Q\left(\mathbf{R}_{\mathrm{u}}\right)\bar{P}\left(\mathbf{R}_{\mathrm{u}}\right)\mathrm{d}\,\mathbf{R}_{\mathrm{u}}\,. \tag{74}$$

This process of integrating out the solvent coordinates is related to an averaging over the solvent states[192,193]

$$\begin{aligned}\bar{P}(\mathbf{R}_{\mathrm{u}})&=\frac{\int\exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{u}},\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}}{\int\exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{u}},\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{u}}\,\mathrm{d}\,\mathbf{R}_{\mathrm{v}}},\\&=\frac{\exp(-\beta W(\mathbf{R}_{\mathrm{u}}))}{\int\exp(-\beta W(\mathbf{R}_{\mathrm{u}}))\,\mathrm{d}\,\mathbf{R}_{\mathrm{u}}}.\end{aligned} \tag{75}$$

resulting in the so-called potential of mean force W, a concept introduced by Kirkwood in the 1930's[191]

$$W(\mathbf{R}_{\mathrm{u}}) = -\beta^{-1}\ln\left(\frac{\int \exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{u}},\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}}{\int \exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}}\right). \tag{76}$$

This potential of mean force (PMF) describes the average solvent structure and is linked to a broad range of properties but, because of the normalization in Eqn. 76, the absolute value of the PMF is meaningless and does not affect the average value in Eqn. 74. Therefore, it is convenient to refer to a system without any solute-solvent interactions, so that the free energy may be expressed as follows[192]

$$\exp(-\beta W(\mathbf{R}_{\mathrm{u}})) = \frac{\int \exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{u}},\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}}{\int \exp\left(-\beta\left(U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{v}}}, \tag{77}$$

where the solvent-solvent terms cancel and the PMF can be rewritten as

$$W(\mathbf{R}_{\mathrm{u}}) = U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right) + \Delta W(\mathbf{R}_{\mathrm{u}}). \tag{78}$$

In this expression, the PMF is divided into the solute potential and a term influenced by the solvent. Introducing a thermodynamic coupling parameter λ for the solute-solvent potential, where λ=0 is the non-interacting reference system and λ =1 the fully interacting system, the solute-solvent interaction part for a given solute geometry can be expressed as

$$\Delta W(\mathbf{R}_{\mathrm{u}}) = \int_0^1 \mathrm{d}\lambda \left\langle \frac{\partial U_{\mathrm{uv}}}{\partial \lambda} \right\rangle_{\mathbf{R}_{\mathrm{u}},\lambda} . \tag{79}$$

This is called coupling parameter integration or (mostly used in the context of MD simulations) thermodynamic integration (TI), where $\langle \ldots \rangle_{\mathbf{R}\mathrm{u},\lambda}$ is the average over the solvent coordinates. The PMF is an effective free energy potential and therefore dependent on the temperature T and the pressure p of the system. A thermodynamic decomposition of the PMF yields

$$\Delta W(\mathbf{R}_{\mathrm{u}}) = \Delta E(\mathbf{R}_{\mathrm{u}}) - T\Delta S(\mathbf{R}_{\mathrm{u}}), \tag{80}$$

with the solvation entropy

$$\Delta S(\mathbf{R}_{\mathrm{u}}) = -\partial \Delta W(\mathbf{R}_{\mathrm{u}})/\partial T . \tag{81}$$

Due to the normalization in Eqn. 76, all these values are excess values with respect to the non-interacting reference system, therefore the excess chemical potential μ^{ex} can be obtained by[192]

$$\exp(-\beta\mu^{ex}) = \frac{\int \exp\left(-\beta\left(U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+\Delta W\left(\mathbf{R}_{\mathrm{u}}\right)\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{u}}}{\int \exp\left(-\beta U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)\right)\mathrm{d}\,\mathbf{R}_{\mathrm{u}}}. \tag{82}$$

The excess chemical potential is the solvation free energy for electronically and structurally frozen solutes, the relaxation of the solute during the solvation process is neglected. To calculate the excess chemical potential using FFs and molecular dynamics simulations, the solute-solvent potential can be split into long (electrostatic) and short range (non-polar) potentials, resulting in a process called free energy decomposition[192]

$$U_{\mathrm{uv}}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)=U_{\mathrm{uv}}^{(\mathrm{np})}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{uv}}^{(\mathrm{elec})}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}\right), \tag{83}$$

with the total PMF

$$W(\mathbf{R}_{\mathrm{u}})=U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+\Delta W^{(\mathrm{np})}(\mathbf{R}_{\mathrm{u}})+\Delta W^{(\mathrm{elec})}(\mathbf{R}_{\mathrm{u}}). \tag{84}$$

Calculation of the potential from only the non-polar PMF is often called cavity formation free energy, calculating it from only the electrostatic PMF charging free energy. They are both related to the last term in Eqn. 65, the non-polar interactions to the Lennard-Jones potential between solute and solvent atoms, the electrostatic interactions to the respective Coulomb potential. If not a single coupling parameter for the solute-solvent potential is introduced, as for the calculation in Eqn. 79, but independent ones for the different parts of this potential, the total potential is given by[192]

$$U\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}:\lambda_1,\lambda_2\right)=U_{\mathrm{u}}\left(\mathbf{R}_{\mathrm{u}}\right)+U_{\mathrm{vv}}\left(\mathbf{R}_{\mathrm{v}}\right)+U_{\mathrm{uv}}^{(\mathrm{np})}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}:\lambda_1\right)+U_{\mathrm{uv}}^{(\mathrm{elec})}\left(\mathbf{R}_{\mathrm{v}},\mathbf{R}_{\mathrm{u}}:\lambda_2\right). \tag{85}$$

From this, the excess chemical potential can be calculated as the reversible work from both contributions using molecular dynamics simulations. The calculation is not path independent, because it depends on the order in which the coupling parameter integrations are performed, usually the cavity formation is the first step in such a calculation

$$\Delta W^{(\mathrm{np})}(\mathbf{R}_{\mathrm{u}})=\int_0^1 \mathrm{d}\lambda_1 \left\langle \frac{\partial U_{\mathrm{uv}}^{(\mathrm{np})}}{\partial \lambda_1} \right\rangle_{\mathbf{R}_{\mathrm{u}},\lambda_1,\lambda_2=0}, \tag{86}$$

followed by the calculation of the charging free energy

$$\Delta W^{(\mathrm{elec})}(\mathbf{R}_{\mathrm{u}})=\int_0^1 \mathrm{d}\lambda_2 \left\langle \frac{\partial U_{\mathrm{uv}}^{(\mathrm{elec})}}{\partial \lambda_2} \right\rangle_{\mathbf{R}_{\mathrm{u}},\lambda_1=1,\lambda_2}. \tag{87}$$

This theory is, as mentioned above, valid for frozen structures. To obtain the solvation free energy, a further averaging over the solute coordinates is needed and can be done by using molecular dynamics simulations. In this work, TI simulations are done using the NAMD 2.11 software.[194] These calculations cause a high computational effort. There are different approaches for TI simulations. The use of frozen solutes results in the excess chemical potential; this approach is comparable to QC solvation models without solute relaxation. Simulations with flexible solutes give access to the solvation free energy, though at the prize of neglected electronic polarization effects with typical fixed-charge force fields.

2.4.2 Dielectric continuum models

In dielectric continuum methods, the solvent is described by a continuum with a specific dielectric constant; the solute is usually embedded in a cavity in this continuum. There are continuum models for QC applications as well as for classical mechanics approaches. The cavity formation free energy can be calculated using a variety of different approaches; often, the cavity is approximated as a sum of spheres with the size of the van der Waals volume of the atoms within the molecule. Detailed discussions about techniques for the cavity formation, like scaled-particle theory, the solvent exposed area and the multipole expansion, can be found in Ref. 192 and 195. The basic ansatz of the solvent exposed area method uses the total surface A of the solute and a constant γ acting as a surface tension between solvent and cavity[193]

$$\Delta W^{(\mathrm{np})}(\mathbf{R}_\mathrm{u}) = \gamma_\mathrm{v} A_\mathrm{tot}(\mathbf{R}_\mathrm{u})\,. \tag{88}$$

The main part in dielectric continuum models is the calculation of the electrostatic contributions to the solvation free energy. With explicit solvent atoms, the electrostatic contribution is given by[192]

$$\Delta W^{(\mathrm{elec})}(\mathbf{R}_\mathrm{u}) = \int_0^1 \mathrm{d}\lambda_2 \left\langle \sum_{i,j} \frac{q_{\mathrm{u},i} q_{\mathrm{v},j}}{\left|\mathbf{r}_{\mathrm{u},i} \text{-} \mathbf{r}_{\mathrm{v},j}\right|} \right\rangle_{\mathbf{R}_\mathrm{u},\lambda_2} , \tag{89}$$

where $\mathbf{r}_{\mathrm{u},i}$ and $\mathbf{r}_{\mathrm{v},j}$ are the position of the solute atom i and solvent atom j. For a continuum solvent, this equation changes to a surface integral

$$\Delta W^{(\mathrm{elec})}(\mathbf{R}_\mathrm{u}) = \frac{1}{2}\int_S d\mathbf{s}' \sum_i q_{\mathrm{u},i} \frac{\sigma(\mathbf{r}')}{\left|\mathbf{r}_{\mathrm{u},i} - \mathbf{r}'\right|} , \tag{90}$$

with the cavity surface **s** and the surface charge density σ over the surface coordinates $\mathbf{r}'$.[192] Often, a Gaussian smearing is used for the surface charge density to overcome discontinuities.[196] Methods using this concept of surface charges are called apparent surface charge (ASC) methods. The surface charges can be calculated using polarization vectors **P**[197]

$$\sigma_{ij} = -\left(\mathbf{P}_j - \mathbf{P}_i\right) \cdot \mathbf{n}_{ij}, \tag{91}$$

where **n** is the unit vector at the surface, pointing from medium *i* to *j*, and the polarization vectors are given using the gradient of the electrostatic potential **V** and the dielectric constant of the medium *ε*

$$\mathbf{P}_i\left(\mathbf{r}\right) = -\frac{\varepsilon_i - 1}{4\pi}\Delta\mathbf{V}\left(\mathbf{r}\right). \tag{92}$$

The electrostatic potential is given by the Poisson equation, where *ρ* is the charge density

$$\nabla\left[\varepsilon\left(\mathbf{r}\right)\nabla\mathbf{V}\left(\mathbf{r}\right)\right] = -4\pi\rho\left(\mathbf{r}\right). \tag{93}$$

If mobile ions are part of the solvent, for example counter ions for charged solutes, the Poisson equation can be expanded to the Poisson-Boltzmann equation by inclusion of the number density of the ions *i* and the PMF between the ions and the solute *w*

$$\nabla\left[\varepsilon\left(\mathbf{r}\right)\nabla\mathbf{V}\left(\mathbf{r}\right)\right] = -4\pi\rho_{\mathrm{u}}\left(\mathbf{r}\right) - 4\pi\sum_i q_i\bar{\rho}_i \exp\left(-\beta w_i\left(\mathbf{r};\mathbf{R}_{\mathrm{u}}\right)\right). \tag{94}$$

The coupling of a continuum solvation model with quantum chemical methods needs the expansion of the Hamilton operator which is given by[197]

$$\hat{H}_{\mathrm{eff}} = \hat{H}_{\mathrm{M}}^{0} + \hat{V}^{\mathrm{int}}, \tag{95}$$

where $\hat{H}_{\mathrm{M}}^{0}$ is the original Hamiltonian of the focused model and $\hat{V}^{\mathrm{int}}$ the solute-solvent interaction operator, which depends on the electrostatic model used. The electrostatic potential is divided into a part inside, $\mathbf{V}_u$, and outside, $\mathbf{V}_v$, of the cavity

$$\mathbf{V}\left(\mathbf{r}\right) = \mathbf{V}_u + \mathbf{V}_v, \tag{96}$$

$$\nabla^2\mathbf{V}_u\left(\mathbf{r}\right) = -\rho\left(\mathbf{r}\right), \tag{97}$$

$$\varepsilon\nabla^2\mathbf{V}_v\left(\mathbf{r}\right) = 0, \tag{98}$$

each with different operators for the respective ASC formulations. There are several ASC methods like the original formulation of the polarizable continuum model (PCM) nowadays called DPCM,[198] CPCM,[199] IEFPCM[200] or COSMO,[201] the details of the respective formulations are shown in the given literature. In this work, the IEFPCM formalism is used as implemented in the Gaussian program package.[154] ASC methods are widespread and can be used in combination with different quantum chemical methods like DFT, HF or MP*n* and can also be used for NMR parameter calculations using GIAOs.[202] Other classes of continuum electrostatic models like multipole expansion (MPE), finite element (FE), and finite difference (FD) methods, need only minor changes in the formalism.[195,197]

The Born models, which were originally developed for spherical ions, like generalized Born (GB) or another method used in this work, the analytical linearized Poisson-Boltzmann（ALPB）method, differ more from ASC approaches and are usually used for molecular mechanics approaches. ALPB is using the Poisson-Boltzmann equation to calculate the so-called electrostatic size of a molecule in an extended version of the Born model; it is in detail discussed in Ref. 203 and 204. While being fast and efficient, dielectric continuum models model only the effect of the solvent onto the solute, the solvent structure is unknown due to the continuum.

2.4.3 Statistical solvation models / Density functional theory

The calculation of the reversible work in Eqn. 79 is associated with averages over the solvent configurations, these configurations can be calculated for example by MD simulations. Assuming that the solute-solvent interaction is composed of the additive contributions of the individual solvent molecules *m* in the configuration $\mathbf{r}_\mathrm{v}$, these averages are given by[192]

$$\begin{aligned}\left\langle \frac{\partial U_{\mathrm{uv}}}{\partial \lambda} \right\rangle_{\mathbf{R}_\mathrm{u},\lambda} &= \left\langle \frac{\partial}{\partial \lambda} \sum_m u_{\mathrm{uv}}\left(\mathbf{R}_\mathrm{u},\mathbf{r}_\mathrm{v};\lambda\right) \right\rangle_{\mathbf{R}_\mathrm{u},\lambda} \\ &= \int d\mathbf{r} \left\langle \sum_m \delta\left(\mathbf{r}\text{-}\mathbf{r}_\mathrm{v}\right) \right\rangle_{\mathbf{R}_\mathrm{u},\lambda} \frac{\partial u_{\mathrm{uv}}\left(\mathbf{R}_\mathrm{u},\mathbf{r};\lambda\right)}{\partial \lambda}. \\ &= \int d\mathbf{r} \left\langle \rho(\mathbf{r}) \right\rangle_{\mathbf{R}_\mathrm{u},\lambda} \frac{\partial u_{\mathrm{uv}}\left(\mathbf{R}_\mathrm{u},\mathbf{r};\lambda\right)}{\partial \lambda}\end{aligned} \tag{99}$$

where δ is the Dirac delta function and $\langle \rho(\mathbf{r}) \rangle_{\mathbf{R}\mathrm{u},\lambda}$ the average solvent density for finding a solvent molecule in configuration $\mathbf{r}_\mathrm{v}$ around the solute configuration $\mathbf{R}_\mathrm{u}$. Calculation of the average solvent density and therefore calculation of the PMF and the excess chemical potential is of great importance for statistical solvation models. The basis for statistical solvation models is the classical density functional theory, which should not be confused with the quantum chemical DFT discussed in chapter *2.2.4*, but has many analogies to it. While in QC DFT the ground state energy is expressed as a functional of the electron density, in classical DFT the thermodynamic properties of the liquid in a specific thermodynamic ensemble are expressed as the functional of the particle density in a simple liquid (liquid of spherical particles, to be extended to molecular liquids below). In the grand canonical ensemble, where the temperature *T*, volume *V* and chemical potential μ, the differential of the energy with respect to the particle number, of the system are given, the grand potential is defined as[205]

$$\Omega = A - \mu N = A - \mu \int \rho(\mathbf{r}) d\mathbf{r}\,, \tag{100}$$

with the Helmholtz free energy *A*, particle number *N* and the density ρ. Influenced by an external potential $\phi(\mathbf{r})$, for example the presence of a solute molecule in the solution, the grand potential changes to

$$\Omega = F + \int \rho^{(1)}(\mathbf{r})\varphi(\mathbf{r}) - \mu N\,, \tag{101}$$

with the single particle density $\rho^{(1)}$ (in the following part of this chapter, the single particle density, which is equivalent to the average solvent density in Eqn. 99, is written ρ, while higher order densities are denoted explicitly). The particle densities can be calculated by

$$\rho^{(n)}\left(\mathbf{r}^{n}\right) = \frac{1}{\Xi}\sum_{N=n}^{\infty}\frac{1}{(N-n)!}\int \exp\left(-\beta V_N\right)\left(\prod_{i=1}^{N} z\exp\left[-\beta\varphi\left(\mathbf{r}_i\right)\right]\right)d\mathbf{r}^{(N-n)}\,, \tag{102}$$

with the activity $z = \exp(\beta\mu)/\Lambda^3$ ($\Lambda = \sqrt{2\pi\beta\hbar^2/m}$ is the thermal wavelength with the mass *m*) and grand partition function

$$\Xi = \sum_{N=0}^{\infty}\frac{1}{N!}\int \exp\left(-\beta V_N\right)\left(\prod_{i=1}^{N} z\exp\left[-\beta\varphi\left(\mathbf{r}_i\right)\right]\right)d\mathbf{r}^{N}\,. \tag{103}$$

The Helmholtz free energy in presence of the external field, *F*, is given by

$$F = A - \int \rho^{(1)}(\mathbf{r})\varphi(\mathbf{r})\,, \tag{104}$$

and can, analogous to the Hohenberg-Kohn theorem, be written as functional of the single particle density

$$F[\rho] = F^{\text{id}}[\rho] + F^{\text{ex}}[\rho]\,, \tag{105}$$

with the ideal part

$$F^{\text{id}}[\rho] = -\beta^{-1}\int \rho^{(1)}(\mathbf{r})\left(\ln\left[\Lambda^3\rho^{(1)}(\mathbf{r})\right] - 1\right)d\mathbf{r}\,. \tag{106}$$

The density minimizing the grand potential is the equilibrium density. The functional for the excess part can be expressed in terms of the equilibrium density ρ_0, via a Taylor series[206]

$$F^{\text{ex}}[\rho] = F^{\text{ex}}[\rho_0] + \int \frac{\delta F^{\text{ex}}[\rho_0]}{\delta\rho(\mathbf{r})}\Delta\rho(\mathbf{r})dr + \frac{1}{2}\int \frac{\delta^2 F^{\text{ex}}[\rho_0]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')}\Delta\rho(\mathbf{r})\Delta\rho(\mathbf{r}')d\mathbf{r}d\mathbf{r}' + \ldots\,, \tag{107}$$

with $\Delta\rho$=ρ-ρ_0 or, using direct correlation functions $c^{(n)}(\mathbf{r}^n)$ and the equilibrium density[205]

$$F^{\text{ex}}[\rho]=F^{\text{ex}}[\rho_0]-\beta^{-1}\int c^{(1)}(\mathbf{r})\Delta\rho(\mathbf{r})d\mathbf{r}-\frac{1}{2}\beta^{-1}\int c^{(2)}(\mathbf{r},\mathbf{r}')\Delta\rho(\mathbf{r})\Delta\rho(\mathbf{r}')d\mathbf{r}d\mathbf{r}'-.... \tag{108}$$

For a uniform fluid, no higher orders than the second order direct correlation function are needed.[206] These direct correlation functions are defined as

$$c^{(1)}(\mathbf{r})=\beta^{-1}\frac{\delta F^{\text{ex}}\left[\rho^{(1)}(\mathbf{r})\right]}{\delta\rho^{(1)}(\mathbf{r})}, \tag{109}$$

respectively,

$$c^{(2)}(\mathbf{r},\mathbf{r}')=\frac{\delta c^{(1)}(\mathbf{r})}{\delta\rho^{(1)}(\mathbf{r}')}=\beta^{-1}\frac{\delta^2 F^{\text{ex}}\left[\rho^{(1)}(\mathbf{r})\right]}{\delta\rho^{(1)}(\mathbf{r})\delta\rho^{(1)}(\mathbf{r}')} \tag{110}$$

with higher order functions $c^{(n+1)}(\mathbf{r}^{n+1})$ being the functional derivatives of $c^{(n)}(\mathbf{r}^n)$. These direct correlation functions can be interpreted as the part of the particle correlations which is not mediated by other particles. Defining a so-called intrinsic chemical potential ψ, which is the functional derivative of the free energy, as the part of the chemical potential not depending on the external potential

$$\frac{\delta F}{\delta\rho(\mathbf{r})}=\psi(\mathbf{r})=\mu-\varphi(\mathbf{r}), \tag{111}$$

as well as calculating the functional derivative of the ideal part of the free energy

$$\frac{\delta F^{\text{id}}}{\delta\rho(\mathbf{r})}=-\beta^{-1}\ln\left[\Lambda^3\rho(\mathbf{r})\right], \tag{112}$$

yields, together with the first order direct correlation function

$$\beta\psi(\mathbf{r})=\beta\frac{\delta F}{\delta\rho(\mathbf{r})}=\ln\left[\Lambda^3\rho(\mathbf{r})\right]-c^{(1)}(\mathbf{r}). \tag{113}$$

This formula shows some interesting aspects: Inserting the definitions of the activity and the intrinsic chemical potential, it can be seen that -$\beta^{-1}c^{(1)}$ is the excess part of the intrinsic chemical potential and therefore the excess chemical potential of the system in absence of the external field. The intrinsic chemical potential is also linked to the density-density correlation function $H^{(n)}$, which is the linear combination of all particle densities including $\rho^{(\text{n})}$ via[205]

$$H^{(2)}\left(\mathbf{r},\mathbf{r}'\right)=-\beta^{-1}\frac{\delta\rho(\mathbf{r})}{\delta\psi(\mathbf{r}')}=\rho\left(\mathbf{r}\right)\rho\left(\mathbf{r}'\right)h^{2}\left(\mathbf{r},\mathbf{r}'\right)+\rho\left(\mathbf{r}\right)\delta\left(\mathbf{r}-\mathbf{r}'\right), \tag{114}$$

with the pair correlation function $h^{(2)}(\mathbf{r}_1,\mathbf{r}_2)=g^{(2)}(\mathbf{r}_1,\mathbf{r}_2)-1$, given by

$$g^{n}\left(\mathbf{r}^{n}\right)=\frac{\rho^{n}(\mathbf{r}_1,\ldots,\mathbf{r}_n)}{\prod_{i=1}^{n}\rho(\mathbf{r}_i)}. \tag{115}$$

The total correlation function is closely related to the PMF and the excess chemical potential. The functional derivative of the intrinsic chemical potential in Eqn. 113 is the inverse function of this density-density correlation function and contains the second order direct correlation function instead of the total correlation function

$$H^{(2)-1}\left(\mathbf{r},\mathbf{r}'\right)=-\beta^{-1}\frac{\delta\psi(\mathbf{r})}{\delta\rho(\mathbf{r}')}=\frac{1}{\rho\left(\mathbf{r}\right)}\delta\left(\mathbf{r}-\mathbf{r}'\right)-c^{(2)}\left(\mathbf{r},\mathbf{r}'\right). \tag{116}$$

Using the functional definition of the delta function with the density-density correlation and its inverse

$$\delta\left(\mathbf{r}-\mathbf{r}'\right)=\int H^{(2)}\left(\mathbf{r},\mathbf{r}''\right)H^{(2)-1}\left(\mathbf{r}'',\mathbf{r}'\right)d\mathbf{r}''=\frac{\delta\rho\left(\mathbf{r}\right)}{\delta\rho\left(\mathbf{r}'\right)}, \tag{117}$$

inserting the functions and performing the integration over $\mathbf{r}''$ yields the Ornstein-Zernike equation

$$h^{(2)}\left(\mathbf{r},\mathbf{r}'\right)=c^{(2)}\left(\mathbf{r},\mathbf{r}'\right)+\int c^{(2)}\left(\mathbf{r},\mathbf{r}''\right)\rho\left(\mathbf{r}''\right)h^{(2)}\left(\mathbf{r}'',\mathbf{r}'\right)\mathrm{d}\mathbf{r}''; \tag{118}$$

respectively, for a uniform and isotropic fluid

$$h^{(2)}\left(r\right)=c^{(2)}\left(r\right)+\rho\int h^{(2)}\left(r'\right)c^{(2)}\left(\left|\mathbf{r}-\mathbf{r}'\right|\right)\mathrm{d}\mathbf{r}'. \tag{119}$$

The Ornstein-Zernike relation connects the total and the direct correlation function. It can be solved recursively, resulting in an infinite series, which can be expressed diagrammatically, showing that the total correlation between the particles at position **r** and **r**′ is given by the direct correlation between them and the indirect correlations via all other, intermediate, particles in the system.

If the overall potential V_N in the grand partition function (Eqn. 103) can be written as the sum of pair potentials u, the excess Helmholtz free energy difference to a reference system with energy F_0^{ex} can be expressed in terms of the pair density and can be calculated, similar to Eqn. 99, via coupling parameter integration. Therefore, the full potential $u(\mathbf{r},\mathbf{r}')=u_0(\mathbf{r},\mathbf{r}')+w(\mathbf{r},\mathbf{r}')$ has to be split into the part of the reference system u_0 and the difference to the full potential w which has to be scaled during the calculation[205]

$$F^{\mathrm{ex}}[\rho] = F_0^{\mathrm{ex}}[\rho] + \frac{1}{2}\int_0^1 d\lambda \iint \rho^{(2)}(\mathbf{r},\mathbf{r}';\lambda)\, w(\mathbf{r},\mathbf{r}')\, d\mathbf{r}d\mathbf{r}'. \tag{120}$$

Using a uniform fluid of density ρ_0 and chemical potential μ_0 as reference system, this becomes

$$\begin{aligned} F^{\mathrm{ex}}[\rho] &= F_0^{\mathrm{ex}} + \mu_0^{\mathrm{ex}} \int \Delta\rho(\mathbf{r})\, d\mathbf{r} - \beta^{-1}\int_0^1 d\lambda (1-\lambda) \iint \Delta\rho(\mathbf{r})\, c^{(2)}(\mathbf{r},\mathbf{r}';\lambda)\, \Delta\rho(\mathbf{r}')\, d\mathbf{r}d\mathbf{r}' \\ &\approx F_0^{\mathrm{ex}} + \mu_0^{\mathrm{ex}} \int \Delta\rho(\mathbf{r})\, d\mathbf{r} - \frac{1}{2}\beta^{-1} \iint \Delta\rho(\mathbf{r})\, c_0^{(2)}(\mathbf{r},\mathbf{r}')\, \Delta\rho(\mathbf{r}')\, d\mathbf{r}d\mathbf{r}' \end{aligned}, \tag{121}$$

where the approximation in the latter part is correct up to the second order. Inserting this approximation and the ideal part given in Eqn. 106 in the grand potential influenced by an external potential (Eqn. 101), the density minimizing this potential is given by

$$\rho(\mathbf{r}) = \rho_0 \exp\left(-\beta\varphi(\mathbf{r}) + \int \rho(\mathbf{r}')\, c_0^{(2)}(|\mathbf{r}-\mathbf{r}'|)\right). \tag{122}$$

This density can be, using the so-called Percus trick, seen as the expression for the pair distribution function of the uniform fluid with density ρ_0 where $\phi(\mathbf{r})$ is the pair potential u

$$g(\mathbf{r}) = \exp\left(-\beta u(\mathbf{r}) + \rho\int c(|\mathbf{r}-\mathbf{r}'|)\, h(\mathbf{r}')\, d\mathbf{r}'\right) = \exp\left(-\beta u(\mathbf{r}) + h(\mathbf{r}) - c(\mathbf{r})\right). \tag{123}$$

The idea of Percus here is to assume that one specific solvent particle can be seen as the solute. This approximation, correct up to the second order, is called hypernetted-chain (HNC) approximation.[207] Higher order correlations are incorporated by the bridge function. The HNC is a so-called closure approximation, since it can be used to solve the Ornstein-Zernike equation in a closed form. There are several closures and approximations which can be used to solve the Ornstein-Zernike equation, these methods can be reviewed in Ref. 205 and the literature given in the respective chapters. The approximation used in this work is based on site-site interactions and is discussed in the next chapter. By calculating the distribution functions, the information about the solvent structure is preserved using statistical solvation models.

2.4.4 The reference interaction site model

The Ornstein-Zernike equation in the given form (Eqn. 118) is valid for simple liquids; nonspherical and anisotropic liquids require incorporation of the angular dependence of the correlation functions. The Ornstein-Zernike equation for such liquids is given by[208]

$$\begin{aligned} h^{(2)}\left(\mathbf{r},\mathbf{\Omega}_{\mathbf{r}},\mathbf{r}',\mathbf{\Omega}_{\mathbf{r}'}\right) &= c^{(2)}\left(\mathbf{r},\mathbf{\Omega}_{\mathbf{r}},\mathbf{r}',\mathbf{\Omega}_{\mathbf{r}'}\right) \\ &\quad +\int c^{(2)}\left(\mathbf{r},\mathbf{\Omega}_{\mathbf{r}},\mathbf{r}'',\mathbf{\Omega}_{\mathbf{r}''}\right)\rho\left(\mathbf{r}'',\mathbf{\Omega}_{\mathbf{r}''}\right)h^{(2)}\left(\mathbf{r}'',\mathbf{\Omega}_{\mathbf{r}''},\mathbf{r}',\mathbf{\Omega}_{\mathbf{r}'}\right)\mathrm{d}\mathbf{r}''\mathrm{d}\mathbf{\Omega}_{\mathbf{r}''}, \end{aligned} \tag{124}$$

with $\mathbf{\Omega}$ showing the rotational dependence of the relative orientation of the molecules incorporated in the correlation functions via the Euler angles. This is still a simplified notation; the derivation of Eqn. 124, called molecular Ornstein-Zernike equation (MOZ), including all definitions of the translational and rotational invariants, is given in Ref. 208. To circumvent the calculation of the MOZ with all the angular dependencies and therefore to reduce the high dimensionality, the reference interaction site model (RISM) can be used.

In RISM theory,[209,210,211] the molecules are composed of spherical interaction sites and are described by a set of site-site correlation functions, reducing the 6D MOZ to a set of 1D integral equations. It is important to note that the site-site HNC theory introduces a systematic error as higher order correlations are neglected, which has to be accounted for. The general form of the closure relation needed to solve the system of integral equations is given (for the 3D case, the 1D expression is similar) in Eqn. 138. The 1D integral equations are the intramolecular correlation function ω, describing the structure of the (rigid) molecule with interaction sites α and γ in distance l[212]

$$\omega_{\alpha\gamma}\left(r\right)=\frac{\delta\left(r-l_{\alpha\gamma}\right)}{4\pi r_{\alpha\gamma}^{2}}, \tag{125}$$

and the direct and total correlation functions as the sum of the pairwise site-site interactions[213]

$$c_{\alpha\gamma}=\sum_{\alpha}\sum_{\gamma}c_{\alpha\gamma}\left(\left|r_{\alpha}-r_{\gamma}\right|\right). \tag{126}$$

The resulting RISM equation is given in matrix form, with the site-site contributions as matrix elements and with the matrix convolution denoted by *, as

$$\mathbf{h}=\boldsymbol{\omega}*\mathbf{c}*\boldsymbol{\omega}+\boldsymbol{\rho}\boldsymbol{\omega}*\mathbf{c}*\mathbf{h}\,. \tag{127}$$

The RISM approach was first introduced for pure solvents, but the formalism can be used for the description of solute-solvent mixtures, whereby the solute is supposed to be in infinite dilution. The RISM equations for such a solute-solvent mixture are more complex. The diagonal density matrix is $\rho_{\alpha_M\gamma_{M'}}=\delta_{\alpha\gamma}\delta_{MM'}\rho_M$ with the Kronecker delta δ; the index α_M denotes the site α of species M in the mixture, and the correlation matrices consist of the elements $c_{\alpha_M\gamma_{M'}}$, $h_{\alpha_M\gamma_{M'}}$ and $\omega_{\alpha_M\gamma_{M'}}$. The RISM equation is

$$\boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho}. \tag{128}$$

Reorganization of the matrices into block matrices including only the solvent-solvent, solute-solvent or solute-solute correlation functions leads to a system of three integral equations. Further, the solute density can be set to zero in the limit of infinite dilution, and inserting $\mathbf{w}^{u/v} = \boldsymbol{\omega}^{u/v} / \boldsymbol{\rho}^{u/v}$ yields the three equations (for the solvent-solvent equation, a single solvent molecule is considered to be the solute)[213]

$$\mathbf{h}^{vv} = \mathbf{w}^{v} * \mathbf{c}^{vv} * \mathbf{w}^{v} + \mathbf{w}^{v} * \mathbf{c}^{vv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv}\boldsymbol{\rho}^{v}, \tag{129}$$

$$\mathbf{h}^{uv} = \mathbf{w}^{u} * \mathbf{c}^{uv} * \mathbf{w}^{v} + \mathbf{w}^{v} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{vv}\boldsymbol{\rho}^{v}, \tag{130}$$

$$\mathbf{h}^{uu} = \mathbf{w}^{u} * \mathbf{c}^{uu} * \mathbf{w}^{u} + \mathbf{w}^{u} * \mathbf{c}^{uv} * \boldsymbol{\rho}^{v}\mathbf{h}^{uv}. \tag{131}$$

These equations are hierarchical, i.e. the total solvent-solvent correlation function can be used for solving the solute-solvent equation, the total solute-solvent correlation function for solving the solute-solute equation. Therefore, it is convenient to define a solvent-susceptibility function χ, often called density response function

$$\boldsymbol{\chi} = \boldsymbol{\rho}^{v}\boldsymbol{\omega}^{v} + \boldsymbol{\rho}^{v}\mathbf{h}^{vv}\boldsymbol{\rho}^{v}, \tag{132}$$

which can be computed from a solvent-solvent calculation and can be used as an input for a solute solvent calculation.

Using the solvent susceptibility, the expression can be simplified to

$$\mathbf{h}^{uv} = \boldsymbol{\omega}^{u} * \mathbf{c}^{uv} * \left(\boldsymbol{\rho}^{v}\right)^{-1} \boldsymbol{\chi}. \tag{133}$$

With these equations, all observables regarding equilibrium properties in solution can be calculated. Here, the excess chemical potential is of special interest, and is given, dependent on the pair potential, by[214]

$$\mu^{\mathrm{ex}} = \rho \sum_{\alpha\gamma} \int dr \int_{0}^{1} d\lambda \left(h_{\alpha\gamma}(r,\lambda) + 1\right) \frac{\partial u_{\alpha\gamma}(r,\lambda)}{\partial \lambda}. \tag{134}$$

The coupling parameter integration in Eqn. 134 is costly because the numerical integration needs a solution of the RISM equations for each λ step. An advantage in the RISM HNC theory is that the expressions for the excess chemical potential and the excess Helmholtz free energy are given in a closed form, since independence of the path prescribed by λ is enforced due to the possibility to write Eqn. 134 as an exact differential. Performing this path independent integration analytically yields the closure dependent expressions for the excess chemical potential. For the HNC closure, this expression is

$$\mu_{\mathrm{HNC}}^{\mathrm{ex}} = -\beta^{-1}\rho\sum_{\alpha\gamma}\int dr\left(\frac{1}{2}h_{\alpha\gamma}^{2} - c_{\alpha\gamma} - \frac{1}{2}h_{\alpha\gamma}c_{\alpha\gamma}\right). \tag{135}$$

Using the HNC closure is often related with numerical instabilities, due to an uncontrolled growth in the exponent.[212] The linearized version of the HNC closure, called Kovalenko-Hirata closure (KH),[215] increases the numerical stability by linearization of the exponent when the argument is larger than a certain threshold. Another class of closures for which path independence is enforced are the *n*-th order partial series expansion (PSE-*n*) closures[216]

$$h_{\alpha\gamma}(r) = \begin{cases} \exp\left(-\beta u_{\alpha\gamma}(r) + h_{\alpha\gamma}(r) - c_{\alpha\gamma}(r)\right) - 1 & \Leftrightarrow -\beta u_{\alpha\gamma}(r) - c_{\alpha\gamma}(r) \leq 0 \\ \sum_{i}^{n} \dfrac{\left(-\beta u_{\alpha\gamma}(r) + h_{\alpha\gamma}(r) - c_{\alpha\gamma}(r)\right)^{i}}{i!} & \Leftrightarrow -\beta u_{\alpha\gamma}(r) - c_{\alpha\gamma}(r) > 0 \end{cases}, \tag{136}$$

They interpolate between both, the KH and HNC closures with increasing order. An detailed derivation of a way to enforce the path independence and the PSE-*n* closures is given in Ref. 214 and 216. The analytical expression for the excess chemical potential for this kind of closure relations is given by (Θ is the Heaviside step function)

$$\mu_{\mathrm{PSE\text{-}n}}^{\mathrm{ex}} = \mu_{\mathrm{HNC}}^{\mathrm{ex}} - \frac{1}{\beta}\sum_{\alpha\gamma}\rho_{\alpha\gamma,\infty}\int\left(\frac{\Theta\left(h_{\alpha\gamma}(r)\right)\left(h_{\alpha\gamma}(r) - \beta u_{\alpha\gamma}(r) - c_{\alpha\gamma}(r)\right)^{n+1}}{(n+1)!}\right)\mathrm{d}r\,. \tag{137}$$

While being computationally fast, even for large systems, the RISM theory and has some known thermodynamic inconsistencies. Besides, the description of the dielectric constant ($\varepsilon_{\mathrm{RISM}}=1+4\pi\beta\rho\mu^{2}/3$ with the solvent dipole moment μ)[212] differs from experimental results using RISM or the extended XRISM[211,217] formalism. For the calculation of salt solutions, the dielectric constant RISM theory (DRISM) was developed,[218,219] achieving consistency for the dielectric constant by adding an additional term, similar to the bridge function, to the closure.

The RISM equations given here are all valid for the one-dimensional theory (1D RISM) but the theory is also usable for three-dimensional problems. Therefore, all spatial distribution functions have to be replaced by the radial distribution functions, resulting in anisotropic solvent distributions around the infinite diluted solute. In analogy to 1D RISM, a precomputed solvent susceptibility, which can be taken from 1D calculations, is used to solve the solute-solvent equation.[220,221] The three-dimensional closure relation, including the bridge function, is given by

$$h_{\alpha\gamma}(\mathbf{r}) = \exp\left(-\beta u_{\alpha\gamma}(\mathbf{r}) + h_{\alpha\gamma}(\mathbf{r}) - c_{\alpha\gamma}(\mathbf{r}) + B_{\alpha\gamma}\right) - 1\,. \tag{138}$$

The solvent susceptibilities used in this work are all calculated using the HNC approximation; therefore, the several approximations of the bridge function are not discussed here. The pair distribution function g_γ can be calculated from the pair correlation function via g_γ=h_γ+1. The pair potential u between the sites α and γ is given by

$$u_{\alpha\gamma} = u_{\alpha\gamma}^{\mathrm{LJ}} + u_{\alpha\gamma}^{\mathrm{elec}} = \sum_{\alpha\gamma} 4\varepsilon_{\alpha\gamma}\left(\left(\frac{\sigma_{\alpha\gamma}}{\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|}\right)^{6}\right) + \sum_{\alpha\gamma}\frac{q_\alpha q_\gamma}{4\pi\varepsilon_0\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|}\,. \tag{139}$$

The parameters σ and ε are the mixed Lennard-Jones parameters (in this work Lorentz-Berthelot mixing rules[222,223] are applied), and q_α and q_γ the partial charges of the respective solute and solvent site. The 3D RISM equations are usually solved on a large three dimensional grid, using numerical techniques to accelerate the convergence, like the modified direct inversion of iterative subspace (MDIIS) algorithm[224,225] as well as the Ewald summation[226] to speed up the calculation of the electrostatic potential, via a split into the short range (SR) and long range (LR) contribution using the Gauss error function (erf) and their complement (erfc) and a smearing factor κ

$$u_{\alpha\gamma}^{\mathrm{elec}} = u_{\alpha\gamma}^{\mathrm{elec,SR}} + u_{\alpha\gamma}^{\mathrm{elec,LR}}\,, \tag{140}$$

$$u_{\alpha\gamma}^{\mathrm{elec,SR}} = \mathrm{erfc}\left(\kappa\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|\right)\sum_{\alpha\gamma}\frac{q_\alpha q_\gamma}{4\pi\varepsilon_0\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|}\,, \tag{141}$$

$$u_{\alpha\gamma}^{\mathrm{elec,LR}} = \mathrm{erf}\left(\kappa\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|\right)\sum_{\alpha\gamma}\frac{q_\alpha q_\gamma}{4\pi\varepsilon_0\left|\mathbf{r}_\alpha - \mathbf{r}_\gamma\right|}\,. \tag{142}$$

Thereby, the short range contribution can be calculated efficiently in real space and the long range electrostatics in reciprocal space after a Fourier transform; more details about 3D RISM calculations avoiding the calculation of real space electrostatics in total can be found in Ref. 227.

2.5 The Embedded-Cluster Reference Interaction Site Model

2.5.1 Basics

The embedded-cluster reference interaction site model (EC-RISM) is a combination of the 3D RISM solvation model with quantum chemical calculations.[88] There are several alternative approaches, most famous

RISM-SCF,[228,229,230] which are not part of this work. For the calculation of the total free energy in solution, not only the excess chemical potential, obtained from 3D RISM calculations, but also the intramolecular energy of this molecule polarized by the solvent is needed. This polarized intramolecular energy can be incorporated within calculations by using polarizable FFs[231] or quantum chemistry, it is not possible using classical, fixed charge, FFs. In the EC-RISM workflow, this is achieved in a self-consistent manner by an iterative cycle:

I. Calculation of the vacuum wave function of the fixed solute, obtaining the electrostatic potential
II. Calculation of the polarized solvent distribution around the solute using this electrostatic potential
III. Embedding the solute in a cluster of point charges representing the solvent and calculation of the wave function and electrostatic potential
IV. Repetition of steps II and III until convergence

The Gibbs free energy in solution, G_{sol}, in the EC-RISM picture is given by the sum of the polarized intramolecular energy E_{sol} of the fixed solute with coordinates $\{\mathbf{r}\}$ and the excess chemical potential using the polarized solvent:

$$G(\{\mathbf{r}_{\mathrm{u}}\})_{\mathrm{sol}} \approx E_{\mathrm{sol}}\left(\{\mathbf{r}\}\right) + \mu^{\mathrm{ex}}\left(\{\mathbf{r}\}\right). \tag{143}$$

By calculating the wave function of the solute embedded in a point charge cluster, the Hamiltonian is expanded by three additional terms, the charge-nuclei, charge-electron and charge-charge interactions. The energy contribution of these three terms has to be subtracted from the total energy using the whole Hamiltonian to obtain the energy of the polarized solute. In practice, the calculation of the charge-charge interaction energy can be suppressed in quantum chemical codes. The energy contribution of the charge-nuclei and charge-electron interactions, the direct interaction energy of the solute with the charges E_q, is given by

$$E_q = \int \rho_q(\mathbf{r})\varphi(\mathbf{r})\,\mathrm{d}\mathbf{r}\ , \tag{144}$$

with the electrostatic potential ϕ at grid point $\mathbf{r}$. The calculation of this energy contribution can be avoided if the quantum chemical code features the direct energetic evaluation of the energy of a given wave function. The charge density ρ_q at solvent site γ can be calculated from the radial distribution function obtained by the 3D RISM calculation

$$\rho_q\left(\mathbf{r}\right) = \sum_{\gamma} q_\gamma \rho_\gamma g_\gamma\left(\mathbf{r}\right), \tag{145}$$

and is used for the construction of the embedding point charge cluster representing the solvent. The radial distribution, and therefore the charge density, are continuous functions; they are discretized to the three dimensional grid with grid spacing Δ*x*, Δ*y* and Δ*z* and grid cell volume Δ*V*=Δ*x*Δ*y*Δ*z*, resulting in the point charges

$$q(\mathbf{r}_i) = \rho_q(\mathbf{r}_i)\Delta V . \tag{146}$$

In Eqn. 139, the 3D RISM potential using the partial charges of the solute sites *α* is shown. Within the EC-RISM workflow, different approaches to calculate the electrostatic potential (ESP) are possible. One is the use of the partial charges as described above, the other one is the use of the full ESP accessible from the QC calculation. This potential can be used for the calculation of the partial charges or directly for the 3D RISM iterations.[231] Therefore, the electrostatic interaction potential at grid point *γ*, $u_\gamma^{\varphi,\text{elec}}$, based on this QC ESP *ϕ*, has to be approximated

$$u_\gamma^{\varphi,\text{elec}} = q_\gamma \varphi(\mathbf{r}), \tag{147}$$

and the full electrostatic potential at this point is built using the difference between full ESP and the point charge based one $\Delta u_\gamma^{\varphi,\text{elec}}$

$$u_\gamma^{\text{elec}} = u_\gamma^{\text{q,elec,SR}} + \Delta u_\gamma^{\varphi,\text{elec}} + u_\gamma^{\text{q,elec,LR}} . \tag{148}$$

This approach can sometimes lead to divergence of the calculations if the difference between full and point charge based ESP not vanish at the box edges. A way to overcome this was developed by Patrick Kibies with a switching function *s* defined as

$$s(r, r_{\min}, r_{\max}) = \begin{cases} 1 & \forall r < r_{\min} \\ s_0 + s_1 r + s_2 r^2 + s_3 r^3 & \forall r_{\max} \le r \le r_{\min} \\ 0 & \forall r < r_{\min} \end{cases} , \tag{149}$$

and results in a switching zone between *r*max, which is chosen as the distance from the molecule to the side of the box, so that the switching zone is the inscribed sphere of the cubic box, and *r*min=*r*max-2Å. The parameters of the switching function are

$$s_0 = \left(3 r_{\min} r_{\max}^2 - r_{\max}^3\right)\left(r_{\min} - r_{\max}\right)^3 , \tag{150}$$

$$s_1 = -6 r_{\min} r_{\max}\left(r_{\min} - r_{\max}\right)^3 , \tag{151}$$

$$s_2 = 3\left(r_{\min} + r_{\max}\right)\left(r_{\min} - r_{\max}\right)^3 , \tag{152}$$

$$s_3 = -2\left(r_{\min} - r_{\max}\right)^3 . \tag{153}$$

This potential switching results in an error of the excess chemical potential, which is given by

$$\Delta_s \mu^{ex} = \int_V g\left(\mathbf{r}\right)\rho_\gamma \Delta_s u_\gamma^{\varphi,\mathrm{elec}} = \int_V g\left(\mathbf{r}\right)\rho_\gamma \left(1 - s\left(r; r_{\min}, r_{\max}\right)\right)\left(u_\gamma^{\varphi,\mathrm{elec}} - u_\gamma^{q,\mathrm{elec}}\right). \tag{154}$$

The EC-RISM method is versatile and can be used for example for the calculation of tautomer fractions,[91] pK_a values,[232,4,5] even for non-aqueous solvents,[233] partition coefficients,[232,8] and NMR parameters[89,234,7] at ambient and even extreme conditions like high hydrostatic pressure.[234,7] A great advantage of EC-RISM is the consideration of the granularity of the solvent.

2.5.2 Empirical correction terms for EC-RISM – the partial molar volume correction

From 3D RISM, especially coupled with quantum chemistry, free energy differences could be computed with high accuracy during the SAMPL2 challenge, but the molecules in the dataset all have a relatively similar size.[91] There is a well-known artifact in 3D RISM theory resulting in too high absolute values for the hydration free energy.[232,235] To overcome this issue, empirical corrections to the excess chemical potential have been introduced.[232,236,237] A way to parameterize such corrections is to use the infinite dilution partial molar volume, V_m or PMV, which can be calculated from Kirkwood-Buff theory within the 3D RISM framework[238,239] and has an almost linear correlation to the error from 3D RISM with respect to experimental solvation free energies. The infinite dilution partial molar volume of the solute can be calculated from the total or direct correlation function ($V_{m,c}$ or $V_{m,h}$)

$$V_{m,c} = \beta^{-1}\kappa_v \left(1 - \rho_v \sum_{\alpha\gamma} \int c_{\alpha\gamma}(\mathbf{r}) 4\pi r^2 \mathrm{d}\mathbf{r}\right), \tag{155}$$

or

$$V_{m,h} = \beta^{-1}\kappa_v - \int h_{\alpha\gamma}(\mathbf{r}) 4\pi r^2 \mathrm{d}\mathbf{r} , \tag{156}$$

using the isothermal compressibility κ of the solute, which can also be calculated from 1D RISM (in the following written as κ_RISM to distinguish between the 1D RISM and experimental solvent compressibility)

$$-\beta^{-1}\kappa_v = \left(\rho_v \left(1 - \rho_v \sum_{\alpha\gamma} c_{\alpha\gamma}\right)\right)^{-1} . \tag{157}$$

Both ways are equivalent for neutral molecules using the RISM compressibility. The correction term used in this work is developed by Daniel Tomazic and Nicolas Tielker and given in the following form[4]

$$\mu^{ex,corr} = \mu^{ex} + c_{V_m} V_m + c_q q \,, \tag{158}$$

with q being the solute charge. The inequality in PMVs calculated via the h- and c-route for charged systems[240,241] is effectively considered by the c_q parameter. The correction parameters have to be parametrized for each solvent model and level of theory. In this work, SPC/E water at 1 bar and 298.15 K is used in EC-RISM at the MP2/6-311+G(d,p) level of theory, and the resulting correction parameters are c_{V_m} =-0.10251 ±0.0005/kcal mol^{-1} Å^3 and c_q=-15.728±0.181/kcal mol^{-1} e^{-1} (for PMVs calculated via the total correlation function using the 3D RISM compressibility). The charge parameter is also necessary for ions since it is needed to correct for the differences between the physical, experimental solvation process and the artificial one, modeled by 3D-RISM. In the experiment, the solute has to cross the interface between the vacuum- and water phase, while this surface potential is neglected in RISM theory since the solvent is infinite and, thus, has no surface.[242] This surface polarization, whose physical background is in detail described in Ref. 243, 244, 245 and 246, needs to be accounted for via the additional parameter, which has no effect for neutral species and cancels within a specific ionization state. Since the parametrization takes the experimental solvation free energies into account, EC-RISM results using this PMV correction should include terms like ideal contributions implicitly. The parametrization is done using only mono charged ions, but yielded good results also for higher charged species during the SAMPL6 challenge.[4] Using this correction, the absolute hydration free energy can be calculated (using Ben-Naim reference states)

$$\Delta_{\text{solv}} G^0_{\text{EC-RISM}} = E^{\text{sol}} + \mu^{\text{ex,corr}} - E^{\text{vac}} \,. \tag{159}$$

The correction parameters were obtained by adjusting to experimental solvation free energies from the Minnesota solvation database (MNSOL)[247] using the following objective function[232,4]

$$\left\{ c_{V_m}, c_q \right\} = \arg\min \left[\sum_i \left(\Delta_{\text{solv}} G^0_{\exp,i} - \Delta_{\text{solv}} G^0_{\text{calc},i} \right)^2 \right]. \tag{160}$$

2.5.3 The partial molar volume correction for extreme conditions

The incorporation of extreme environmental conditions, like high temperatures, pressures or salt concentrations into solvation models is of great importance for the understanding of the behavior of molecules in biological systems. There are multiple ways to achieve that, explicit solvation in MD or, at a quantum chemical level, ab-initio MD simulations are usable as well as approaches in implicit solvation models, like PCM-XP.[248,249] The 3D RISM theory, and therefore the EC-RISM solvation model, is also able to incorporate the effect of temperature and pressure changes because the solvent density is considered in the 3D RISM equations and the solvent susceptibility, which also includes the pressure dependence of the dielectric constant, when, as in this work, computed with DRISM. While the polarizing effect of the high-pressure solvent is considered using the quantum chemistry-derived partial charges or ESP, the FF used in the 3D RISM calculations remains the same as for ambient condition calculations. Besides, the PMV-correction, which compensates for an error in the cavity formation energy and, thus, should be pressure dependent, is only parametrized for ambient conditions. Therefore, a systematic shift of EC-RISM solvation free energies compared to pressure dependent reference data can be observed. So, to calculate accurate absolute hydration free energies at high pressure, another empirical correction term is needed, this correction was developed and parametrized by N. Tielker.[7] Due to the lack of experimental reference data, suitable values have to be computed:

$$\Delta_{\text{solv}}G^0_{\text{ref}}\left(p\right)=\Delta_{\text{solv}}G^0_{\text{TI}}\left(p\right)+\Delta_{\text{solv}}G^0_{\text{EC-RISM}}\left(1\text{bar}\right)-\Delta_{\text{solv}}G^0_{\text{TI}}\left(1\text{bar}\right). \qquad (161)$$

The reference data for high pressure are the 1 bar corrected EC-RISM values $\Delta_{\text{solv}}G^0_{\text{EC-RISM}}\left(1\text{bar}\right)$ adding the pressure dependence from TI calculations. With these reference data, a pressure dependent correction in the form

$$\mu^{\text{ex,corr,HP}}\left(p\right)=\mu^{\text{ex,corr}}\left(p\right)+c_{\text{HP}}\left(p-1\text{bar}\right)V_m\left(p\right), \qquad (162)$$

can be parametrized. Therefore, the pressure dependent solvation free energy is needed

$$\Delta_{\text{solv}}G^0_{\text{EC-RISM}}\left(p\right)=E^{\text{sol}}\left(p\right)+\mu^{\text{ex,corr,HP}}\left(p\right)-E^{\text{vac}}, \qquad (163)$$

the pressure dependent electronic solute energy is directly obtained from a pressure dependent EC-RISM calculation. The $\mu^{\text{ex,corr}}(p)$ is the excess chemical potential from an pressure dependent EC-RISM calculation by applying the PMV correction for ambient conditions. These reference data are computed for a subset of the MNSOL, which is described in Ref. 7. The objective function for the parametrization is

$$\left\{c_{\text{HP}}\right\}=\arg\min\left[\sum_{i,j}\left(\Delta_{\text{solv}}G^0_{\text{ref},i}\left(p_j\right)-\Delta_{\text{solv}}G^0_{\text{EC-RISM},i}\left(p_j\right)\right)^2\right], \qquad (164)$$

resulting in a high pressure correction parameter of c_{HP}=-1.0108 10^{-5} kcal mol^{-1} Å^{-3} bar^{-1} for

MP2/6-311+G(d,p)/EC-RISM with the solvent model described in Ref. 7. An alternative approach is the calculation of the explicit route within the pressure dependent thermodynamic cycle, this route is given by[7]

$$\begin{aligned}\Delta G_1\,(p_1 \rightarrow p_2) = {} & E_{\mathrm{sol}}(p_2,T) - E_{\mathrm{sol}}(p_1,T) - T(S_{\mathrm{sol}}(p_2,T) - S_{\mathrm{sol}}(p_1,T)) \\ & + \mu_{\mathrm{sol}}^{\mathrm{ex}}(p_2,T) - \mu_{\mathrm{sol}}^{\mathrm{ex}}(p_1,T) + \mu_{\mathrm{sol}}^{\mathrm{id}}(p_2,T) - \mu_{\mathrm{sol}}^{\mathrm{id}}(p_1,T)\end{aligned}, \tag{165}$$

with the ideal part being, as derived by S. Kast

$$\mu_{\mathrm{sol}}^{\mathrm{u,id}}(p_2,T) - \mu_{\mathrm{sol}}^{\mathrm{u,id}}(p_1,T) = \int_{p_1}^{p_2} V_{\mathrm{m}}^{\mathrm{u,id}}(p,T)\,dp = RT\int_{p_1}^{p_2} \kappa_{\mathrm{v}}(p,T)\,dp\,, \tag{166}$$

which makes the calculation of pressure dependent free energies possible via an integration of the pressure dependent PMV. This approach is not used in this work, instead the pressure dependent PMV correction is used. For the temperature correction, for consistency reasons, a similar ansatz is chosen. The reference data for the parametrization also have to be computed using frozen body TI calculations and are given by

$$\Delta_{\mathrm{solv}} G_{\mathrm{ref}}^0(T) = \Delta_{\mathrm{solv}} G_{\mathrm{TI}}^0(T) + (\Delta_{\mathrm{solv}} G_{\text{EC-RISM}}^0(298.15\,\mathrm{K}) - \Delta_{\mathrm{solv}} G_{\mathrm{TI}}^0(298.15\,\mathrm{K}))\,. \tag{167}$$

The corresponding correction term is

$$\mu^{\mathrm{ex,corr,T}}(T) = \mu^{\mathrm{ex,corr}}(T) + c_T\,(T - 298.15\mathrm{K})\,V_m(T)\,, \tag{168}$$

the temperature dependent solvation free energy

$$\Delta_{\mathrm{solv}} G_{\text{EC-RISM}}^0(T) = E^{\mathrm{sol}}(T) + \mu^{\mathrm{ex,corr,T}}(T) - E^{\mathrm{vac}}\,, \tag{169}$$

and the objective function for the parametrization is

$$\{c_T\} = \arg\min\left[\sum_{i,j}\left(\Delta_{\mathrm{solv}} G_{\mathrm{ref},i}^0(T_j) - \Delta_{\mathrm{solv}} G_{\text{EC-RISM},i}^0(T_j)\right)^2\right]. \tag{170}$$

The parametrization of the temperature dependent correction as well as the performance on the trainings set and an independent test set is described in chapters *3.2.1* and 4.1. A detailed derivation of temperature derivatives of observables, like the PMV, within the 3D-RISM framework and an alternative correction scheme are described in Ref. 250.

There are already alternative, even 3D RISM based,[251] approaches to incorporate temperature dependence in computational solvation models using temperature dependent experimental data.[252,253] The experimental dataset from Ref. 252 and 253, here called Chamberlin dataset, incorporates a large variety of molecules and temperatures. The molecules for the test set (a subset of the Chamberlin dataset, given in chapter *3.2.1*) are not very complex, they are small organic compounds, include halogen atoms, and have a small number of rotatable bonds. The last point is needed since no conformational sampling or search for different tautomers is done for the test set. For all of the test set compounds, experimental solvation free energies are given over a broad range of temperatures, but unfortunately these energies are not given for ambient conditions since the data are provided only in form of the minimum and maximum temperature available with the respective experimental values. Temperature dependent experimental reference data are not available in the MNSOL database used for the parametrization of all PMV corrections for EC-RISM. For consistency, the temperature dependent correction is parametrized in a similar fashion as the pressure dependent correction, using the same dataset of molecules and TI reference calculations. This way it is possible to use the existing PMV correction for ambient calculations and train only a single, temperature dependent term to expand this correction to temperature variations. This approach gives computational consistency, retains the original PMV correction for ambient conditions and allows the use of the temperature dependent experimental data of the Chamberlin dataset as a benchmark dataset; the resulting parameter is c_T=-1.65571 10^{-4} kcal mol^{-1} $\mathrm{Å}^{-3}$ K^{-1}.

2.6 Calculation of conformational, tautomer and ionization state fractions and dissociation constants

To understand the behavior of a molecule in solution, a large variety of conformational, tautomeric and ionization states has to be considered. Information about the pH-dependent ionization state fractions and the underlying conformational and tautomeric ensembles is needed to calculate NMR spectra for comparison with experimental ones. To calculate the dissociation constant of an acid base reaction, at first the reaction has to be defined. Here, we define it as a deprotonation reaction of a substance with *n* titratable sites among which *i* are deprotonated, $\mathrm{AH}_{n-i} \rightarrow \mathrm{AH}_{n-i-1} + \mathrm{H}^+$. The equilibrium constant is[4]

$$-\beta^{-1} \ln K_{a,i+1} = -\beta^{-1} \ln \frac{a_+ \, a_{i-1}}{a_i} = \Delta_r G^0 = \mu_+^0 + \mu_{i-1}^0 - \mu_i^0 , \tag{171}$$

with the activity of the proton a_+ and the standard chemical potentials in solution, referenced to an infinite dilution standard state of 1 bar and c^0=1 M at a specified temperature. In this work, the notation for the standard states is neglected for clarity.[254] These chemical potentials can for instance be computed by quantum chemistry and are approximated by[5]

$$\mu_j(i) \approx \mu_j^{\text{id}}(i) + E_j^{\text{sol}}(i) + \mu_j^{\text{ex}}(i) + G_j^{\text{RRHO}}(i) \equiv \mu_j^{\text{id}}(i) + G_j(i), \tag{172}$$

where the subscript *j* denotes a fixed conformational and tautomeric state of molecule *i*, and G_j^{RRHO} is the rigid rotor, harmonic oscillator (RRHO) model, describing the rotational and vibrational contributions to the Gibbs energy. The reference to the standard state is included in the ideal parts, the interaction part G_j can be calculated using EC-RISM theory, where the infinite dilution condition is directly included in the Hamiltonian (therefore, the superscript is left out in the following). Assuming that the activities in Eqn. 171 can be replaced by the molar fractions *x*, the fractions of the states A$^-$ and AH for a molecule with only a single titratable site are given by

$$x_0 = x(\text{AH}) = \frac{1}{1 + a_+^{-1} K_{a,1}}, \tag{173}$$

$$x_1 = x(\text{A}^-) = \frac{a_+^{-1} K_{a,1}}{1 + a_+^{-1} K_{a,1}}, \tag{174}$$

and for a system with two titratable sites AH$_2$

$$x_0 = x(\text{AH}_2) = \frac{1}{1 + a_+^{-1} K_{a,1} + a_+^{-2} K_{a,1} K_{a,2}}, \tag{175}$$

$$x_1 = x(\text{AH}^-) = \frac{a_+^{-1} K_{a,1}}{1 + a_+^{-1} K_{a,1} + a_+^{-2} K_{a,1} K_{a,2}}, \tag{176}$$

$$x_2 = x(\text{A}^{2-}) = \frac{a_+^{-2} K_{a,1} K_{a,2}}{1 + a_+^{-1} K_{a,1} + a_+^{-2} K_{a,1} K_{a,2}}. \tag{177}$$

Further, replacing the activity of the proton by $10^{-\text{pH}}$, the equation can be generalized to calculate the pH-dependent fractions of each ionization state in a system of *n* titratable sites[4]

$$x_i = (10^{-\text{pH}})^{-i} \prod_{j=0}^{i} 10^{-\text{p}K\text{a},j} \left(1 + \sum_{k=1}^{n} (10^{-\text{pH}})^{-k} \prod_{l=1}^{k} 10^{-\text{p}K\text{a},l} \right)^{-1}. \tag{178}$$

The system may contain multiple underlying tautomers, each with underlying conformational degrees of freedom, using the discrete partition function formalism. The conditional population of a conformation *c* within a tautomer state *t* of a given ionization state *i* is given by

$$x_{itc|it} = \exp(-\beta G_{itc|it}) / Z_{it|i}, \tag{179}$$

with the partition function

$$Z_{it|i} = \sum_{c} \exp(-\beta G_{itc|it}), \tag{180}$$

which is related to the energy of the specific tautomer. The latter is given using the predominant states approximation[255] as (neglecting explicitly considering RRHO contributions since they are implicitly included in the PMV-correction parameters)

$$G_{it|i} = -\beta^{-1} \ln(\int d\mathbf{R}_{itc|it} \exp(-\beta G_{itc|it}(\mathbf{R}_{itc|it}))) \approx -\beta^{-1} \ln(\sum_{itc|it} \exp(-\beta G_{itc|it})). \tag{181}$$

Similarly, the populations of a tautomer within a respective ionization state, the partition function, and the energy of this ionization state are

$$x_{it|i} = \exp(-\beta G_{it|i}) / Z_i, \tag{182}$$

$$Z_i = \sum_{t} \exp(-\beta G_{it|i}), \tag{183}$$

$$G_i = -\beta^{-1} \ln(\sum_{it|i} \exp(-\beta G_{it|i})). \tag{184}$$

The pH-dependent tautomeric and conformational fractions can then be calculated by multiplication of the respective fractions $x_{itc} = x_{itc|it} x_{it|i} x_i$. By combining this approach to treat such multistate species with Eqn. 171 and 172, and neglecting the pressure-volume contributions, as described in Ref. 5, the reaction free energy is given using a summation over acid, N, and base states, M

$$\begin{aligned} \Delta_{\mathrm{r}} G^0 &= \mu^{\mathrm{id}}(\mathrm{H}^+_{\mathrm{aq}}) + \mu^{\mathrm{id}}(\mathrm{A}^-_{\mathrm{aq}}) - \mu^{\mathrm{id}}(\mathrm{HA}_{\mathrm{aq}}) \\ &+ \mu^{\mathrm{ex}}(\mathrm{H}^+_{\mathrm{aq}}) - \beta^{-1} \ln \frac{\sum_{j=1}^{M} \exp[-\beta G_j(\mathrm{A}^-_{\mathrm{aq}})]}{\sum_{k=1}^{N} \exp[-\beta G_k(\mathrm{HA}_{\mathrm{aq}})]} \;. \end{aligned} \tag{185}$$

The fifth term of this equation can be calculated using standard solvation models, the first three are ideal contributions and the fourth is the Gibbs free energy of hydration of the proton, $\mathrm{H}^+_{\mathrm{aq}}$.[256,257] Assuming the first four terms to be an additive constant and switching to the decadic pK scale results in[5]

$$\mathrm{p}K_{\mathrm{a}} = \frac{\beta \Delta_{\mathrm{r}} G^0}{\ln 10} = b - \frac{m}{\ln 10} \ln \frac{\sum_{j=1}^{M} \exp[-\beta G_j(\mathrm{A}^-_{\mathrm{aq}})]}{\sum_{k=1}^{N} \exp[-\beta G_k(\mathrm{HA}_{\mathrm{aq}})]}, \tag{186}$$

with the constant b which does not need to be computed because it can be adjusted by fitting to experimental reference data

$$b = \frac{\beta}{\ln 10}[\mu^{0,\mathrm{id}}(\mathrm{H}^+_{\mathrm{aq}}) + \mu^{0,\mathrm{id}}(\mathrm{A}^-_{\mathrm{aq}}) - \mu^{0,\mathrm{id}}(\mathrm{HA}_{\mathrm{aq}}) + \mu^{\mathrm{ex}}(\mathrm{H}^+_{\mathrm{aq}})]\,. \tag{187}$$

In this approach, there is another parameter, *m*, which would ideally be 1 but also can be adjusted like *b* because it introduces more computational flexibility and because a lot of reference data is available.[232,4,233] The parametrization used in this work was done by Nicolas Tielker[4] using the dataset of Klicić *et al.*[258] resulting in regression parameters *m*=0.74±0.02 and *b*=-150.72±4.11 at the MP2/6-311+G(d,p)/EC-RISM(PSE-2) level of theory. This pK_a model and level of theory was successfully used during the SAMPL6 and SAMPL7 challenges, but it was not available during the time of the SAMPL5 challenge. This made recalculations of the SAMPL5 dataset using this level of theory necessary to judge the improvements made during the SAMPL6 and 7 challenges. The results of the new theoretical developments applied on older SAMPL datasets are reviewed in Ref. 10.

III. Methods

3.1 General computational details

If not stated otherwise, the calculations in this work were performed similar to the approach developed for the SAMPL6 pK_a prediction challenge.[4] This approach is described in the following sections. Deviations from this approach for each specific chapter are given in the last section of this passage.

3.1.1 Geometry optimizations and single point calculations

The geometry optimizations were performed at the B3LYP/6-311+G(d,p)/PCM[131,132,197] level of theory using default settings and tight convergence criteria in Gaussian 09 rev E0.1[154] or Gaussian 16 rev B0.1[259] (the actual version used is given for the respective chapters), yielding the solution (PCM) structures which are used for EC-RISM and MP2/6-311+G(d,p)/PCM single point calculations as well as TI calculations. The solution structures were reoptimized at the B3LYP/6-311+G(d,p) level using the same Gaussian version used for the solution structure including a frequency calculation allowing the calculation of thermal corrections. These structures were used for the vacuum calculations, MP2/6-311+G(d,p), again using the same Gaussian version and CCSD(T)/cc-pVTZ calculations with Orca 4.0.1[130] using the RI/F12 approximations.[119,129]

3.1.2 EC-RISM calculations

The EC-RISM calculations were conducted at the 6-311+G(d,p)/PSE-2 level of theory in Gaussian 09 rev E0.1[154] using the exact Hartree-Fock electrostatic potentials during iterations and MP2 energy calculations for the final iteration. The 3D RISM calculations within EC-RISM were performed on a cubic grid (140^3 grid points with a 0.3 Å spacing), using the solvent susceptibilities from Ref. 91 if only ambient conditions are considered, and Ref. 7 for high pressure calculations. The generation of the solvent susceptibilities for temperature variations are described in chapter *3.1.3*. These susceptibilities were generated with the dielectrically consistent DRISM/HNC [218,219] model and a modified version of the SPC/E water model[260,261] using a temperature of 298.15 K, a bulk density of 0.03334 Å^{-3} and a dielectric constant of 78 for ambient conditions and the values from Floriano et al.[262] for the high pressure calculations, respectively; these values are given in Table 1. The MDIIS algorithm[224] was used to improve the convergence of the RISM equations. The convergence criterion for solvent susceptibility generation was 10^{-8} for the maximum residual of the direct correlation function. The calculations were performed on a logarithmically spaced grid ranging from 0.0059 Å to 164.02 Å.

The solute Lennard-Jones parameters were taken from GAFF version 1.7.[184] Within EC-RISM, 10^{-6} is used as convergence criterion for the maximum residual norm of direct correlation functions for the 3D RISM calculations and 0.01 kcal mol^{-1} for the energy difference between two consecutive EC-RISM iterations. The CHelpG[263] scheme with default radii was used to calculate atom centered point-charges as well as a dipole constraint enforcing the reproduction of the quantum mechanically obtained dipole moment and a point charge compressor.[18] NMR parameter (EC-RISM/GIAO)[7,89,234] calculations were done at the MP2/6-311+G(d,p) level of theory using Gaussian 09 rev E0.1.

For all EC-RISM calculations the κ_{RISM} was used and the partial molar volumes calculated via the *c*-route was used together with the parameters described in chapter *2.5.2* when applying the PMV-correction.

Table 1: Densities ρ, dielectric constants ε and isothermal compressibilities κ of water, used for the calculation of the high-pressure solvent susceptibilities, all at a temperature of 298.15 K. The values were calculated using the equations of state from Floriano et al.[262] Besides, 1D RISM compressibilities are given, data and solvent susceptibilities are taken from Ref. 7.

p/bar	ε	ρ/Å^{-3}	κ/10^9 Pa^{-1}	κ_{RISM}/10^9 Pa^{-1}
1	78.45	0.03333	0.4502	0.7141
100	78.85	0.03348	0.4391	0.7070
500	80.37	0.03404	0.3990	0.6334
1000	82.07	0.03460	0.3576	0.5695
2000	84.95	0.03583	0.2950	0.4750
3000	87.34	0.03682	0.2502	0.4086
4000	89.38	0.03769	0.2165	0.3595
5000	91.17	0.03846	0.1905	0.3217
7500	94.84	0.04009	0.1455	0.2566
10000	97.75	0.04142	0.1170	0.2152

3.1.3 Generation of solvent susceptibilities for various temperatures

The solvent susceptibilities for different temperatures were generated in a similar fashion as the high pressure susceptibilities. The dielectrically consistent DRISM/HNC[218,219] theory and a modified version of the SPC/E water model[260,261] were used, and the temperature, water density and dielectric constants are adjusted with the values taken from equations of state from the official releases of the international association for the properties of water and steam (IAPWS, http://www.iapws.org)[264,265] using the implementation of the NIST Chemistry WebBook (http://www.webbook.nist.gov accessed 12.02.2019). The values used are given in Table 2. The convergence of the RISM equations, which are solved on a logarithmically spaced grid ranging from 0.0059 Å to 164.02 Å, was improved by using the MDIIS algorithm,[224] and the convergence criterion was set to 10^{-8} for the maximum residual of the direct correlation function.

Table 2: Densities ρ, dielectric constants ε and isothermal RISM compressibilities κ_{RISM} of water, used for the calculation of the temperature dependent solvent susceptibilities. The values were calculated using the equations of state from the IAPWS.[264,265]

T/K	ε	ρ/Å^{-3}	$\kappa_{\mathrm{RISM}}/10^9$ Pa^{-1}
278.15	85.9155	0.0334616	0.763150
283.15	83.9745	0.0334528	0.747943
288.15	82.0775	0.0334327	0.734873
293.15	80.2226	0.0334028	0.723616
298.15	78.4084	0.0333640	0.713923
303.15	76.6340	0.0333172	0.705596
308.15	74.8983	0.0332631	0.698478
313.15	73.2005	0.0332023	0.692441
318.15	71.5401	0.0331353	0.687376
323.15	69.9160	0.0330624	0.683209
328.15	69.3277	0.0329840	0.679857
333.15	66.7745	0.0329004	0.677264
338.15	65.2550	0.0328119	0.675380
343.15	63.7703	0.0327187	0.674163
348.15	62.3180	0.0326209	0.673577
363.15	58.1519	0.0323019	0.675340
368.15	56.8245	0.0321874	0.677035
372.756	55.6279	0.0320785	0.679067

While these solvent susceptibilities were generated for the isobaric case, with only temperature variations, it is easily possible to generate solvent susceptibilities for varying temperatures and pressures with the adjustments made in the 1D RISM Mathematica notebook[266] implementation for the generation of these solvent susceptibilities.

It is important to notice that there are three different solvent susceptibilities for ambient conditions, the one generated for ambient conditions (Ref. 91), one in the high pressure (Ref. 7) and one in the various temperature series (described above). The ambient condition solvent susceptibility was generated using the experimental water density, while for the other ones, for consistency reasons within the series of solvent susceptibilities, densities from equations of state are used. In this work, the original, ambient condition solvent susceptibility was only used for the investigation of systems for which only ambient conditions were considered (chapter 4.2 and 4.3); for the pressure- and temperature dependent calculations, the ambient condition solvent susceptibility from the respective series was used.

3.1.4 TI calculations

The simulation boxes were set up by placing 4167 SPC/E water molecules in a cubic box with an edge length of 50 Å around a single solute molecule using the packmol 1.1.2.023 software.[267] The used structures are the equivalent PCM optimized structures, used also for the EC-RISM calculations. AM1-BCC[75,268] charges and GAFF parameters[184] (version 1.7) were used for the simulations which were performed using NAMD 2.11.[194] Within the simulations, the 1-4 interactions (which are irrelevant for frozen structure calculations, frozen structures are enforced with the fixedAtoms and fixedAtomsForces options in NAMD) were scaled by 0.833333 and the water bonds were kept rigid using the SETTLE algorithm,[269] while the Lennard-Jones interactions were gradually switched off between 10 and 12 Å. The electrostatic interactions are calculated efficiently using Particle Mesh Ewald (PME) algorithm in the fourth order with a 1.0 Å grid spacing.[270] The temperatures and pressures are kept constant at the respective values using Langevin dynamics and the Nosé-Hoover-Langevin piston.[271,272] Each system was first minimized for 5000 steps using the conjugate gradient and line search algorithms as implemented in NAMD, followed by an equilibration for 0.4 ns with a 2 fs time step and switched off solute-solvent interactions. The TI coupling parameter λ was increased linearly for the Lennard-Jones interactions between 0 and 1 in steps of 0.1 during the simulation, followed by a similar switching on of the electrostatic interactions and the calculation of the hysteresis in the reverse order. For each λ-step, the system was equilibrated for 60 ps and afterwards simulated for 0.4 ns.

3.2 Computational details of the specific chapters

3.2.1 EC-RISM for temperature variations: Parametrization and benchmarking

The molecular structures used for the parametrization are divided into two datasets, a training and a test set. The training set for the parametrization was taken from Ref. 7. The TI-MD simulations were done considering only the single rigid minimum conformation of each of the 47 molecules without further optimization since they were prepared using the SAMPL6 workflow. The same set of structures was also used for the EC-RISM calculations. All calculations were done using the solvent susceptibilities described in 3.1.3 and the computational workflow outlined in 3.1; the calculations are done at 278.15, 283.15, 288.15, 293.15, 298.15, 303.15, 308.15, 313.15, 318.15, 323.15, 328.15, 333.15, 338.15, 343.15, 348.15 and 363.15 K. Resulting in 846 TI-MD simulations.

The structures of test-set are generated using Avogadro[186] without any conformational or tautomer sampling and optimized analogous to the training-set molecules. The test set contains 27 molecules (1-ethoxybutane, 2-methylpentane, 2-propanol, 3-nitrophenol, benzene, benzonitrile, chloromethane, cyclopropane, dimethylsulfide, ethene, ethylbenzene, ethylbenzoate, fluoromethane, hexane, hydrogensulfide, methane, methylamine, morpholine, octanoic-acid, phenol, piperidine, propanethiol, quinoline, tetrafluoroethylene, tetrahydrofuran, trichloroethylene, urea) which were taken from Ref. 252 and 253; these molecules were chosen because there is experimental reference data for a large available temperature interval in the literature, and they have little conformational freedom. The latter is especially important since conformations and tautomers were neglected during the structure generation of the test set. Calculations were performed as described in chapter 3.1 using Gaussian 16 rev B01. The structures of the test and trainings set molecules are given in the structures subfolder of the SI part 01.

3.2.2 Calibrating tautomer predictions: the SAMPL2 datasets

The conformations of all molecules were taken from Ref. 91 and re-optimized at B3LYP/6-311+G(d,p)/PCM level of theory using the computational approach described in chapter 3.1 using Gaussian 09. The TI, EC-RISM, CCSD(T) and frequency calculations also followed this workflow; NMR parameters were not computed. The raw data for each molecule, the structures, data for all pairs, from the original SAMPL2 submission as well as Ref. 100 are given in the SI part 02.

3.2.3 Probing the sensitivity of chemical shifts for tautomer predictions - histamine

Computational details

An exhaustive conformational sampling was done for all of the histamine tautomers via a relaxed dihedral scan of all three rotatable bonds in steps of 90°, yielding 64 conformations per tautomer. These structures were afterwards submitted to an unconstrained geometry optimization as described in chapter 3.1 using Gaussian 09, and all conformations within a Cartesian RMSD of 0.1 Å were condensed into the minimum conformation. These conformations were used for EC-RISM/NMR calculations. The structures for the reparametrization of the GAFF n3 atomtype Lennard-Jones parameters, which was necessary to improve the accuracy of the histamine pK_a-calculations, were taken from Ref. 4 and can be found in the GAFF_Reparm/Structures subfolder of SI part 03. EC-RISM calculations were performed, again, following the workflow given in chapter 3.1. To calculate the free energy derivatives, λ-steps of 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 were chosen, but only the endpoints are needed to calculate the derivatives.

Experimental details

To get access to representative spectroscopic data for the histamine free base and the monocation, measurements were conducted at different pH values. The pK_a values of histamine are 9.75 and 6.04[273], so pH 12.5 was chosen as representative for the free base with a free base fraction of 0.998, and pH 7.9 for the monocation with a fraction of 0.973. The histamine was purchased from Sigma Aldrich and used without further purification; for each sample approx. 25 µg histamine were dispensed in 600 µL demineralized water, using a Milli-Q site, at 22 °C. The pH values were adjusted using NaOH and HCl.

The NMR spectroscopy was done by Wolf Hiller using TMS in methylene chloride as external reference on a Bruker Avance III HD NMR-spectrometer with 600 MHz and a BBFO Cryo sample holder chilled with helium. The nitrogen shifts were referenced using the spectrometer calibration to nitromethane in heavy water. Direct ^{1}H, ^{13}C and ^{15}N experiments were done together with ^{1}H/^{1}H COSY, ^{1}H/^{13}C HSQC and HMBC as well as ^{1}H/^{15}N HSQC and HMBC measurements.

3.2.3.1 Optimization of force field parameters

When a usual molecular mechanics FF is used, the Lennard-Jones parameters have to be assigned to the respective atoms of the molecule. The simplest way to do this is the assignment according to the elements but this simple assignment is not sufficient for the description of real molecules. The same element can have strongly differing properties within the same molecule depending on the chemical environment, like neighboring atoms or the bonding situation. Therefore, the Lennard-Jones parameters are usually not element dependent and instead assigned according to so-called atom types. This way it is for example easily possible to distinguish between aromatic and aliphatic carbons, carbonyl and hydroxyl oxygen atoms or charged and uncharged nitrogen atoms. Also, it is easy to parametrize specific atomic groups, and even a reparametrization of specific atom types is possible without changing the entire system. In Chapter 4.3 of this work, the neurotransmitter histamine is discussed. During the analysis of the results, it was found that for those molecules in the MNSOL database that contain nitrogen atoms with the atomtype n3, experimental Gibbs free energies of hydration are only reproduced with a large RMSE. The n3 type is simply the atomtype for sp^3 nitrogen atoms with three bonds. Therefore, reparametrization of this atomtype was done based on the physical insight given by 3D RISM calculations. The reparametrization method developed during this work is a refined version of a simpler one used in J. Auch's Bachelor's thesis.[274] It is based on an in-house code developed by Yannic Alber which allows to calculate so-called free energy derivatives. These free energy derivatives were also used in a 3D RISM based approach for the reparametrization of FF parameters in a C. Chodun's Master's thesis[275] and are derivatives of the excess chemical potential with respect to the potential. They are given, for the PSE-closure, by[216,276]

$$\frac{\delta \mu^{\text{PSE-}n}}{\delta u(\mathbf{r})} = \rho_\gamma(\mathbf{r}) g_\gamma(\mathbf{r}), \tag{188}$$

and further, by transforming the variation to a derivative with respect to a Lennard-Jones parameter or the charge (here called θ)

$$\frac{\partial \mu^{\text{PSE-}n}}{\partial \theta} = \rho_\gamma \sum_\gamma \int g_\gamma(\mathbf{r}) \frac{\partial u_{\alpha,\gamma}(\mathbf{r})}{\partial \theta_\alpha} d\mathbf{r}\,. \tag{189}$$

The following approach is possible using these free energy derivatives:

1) Calculation of the Gibbs free energies of hydration for all neutral molecules of the MNSOL database containing the n3 parameter using EC-RISM and the PMV correction.

2) Calculation of the difference between the experimental and EC-RISM results.
3) Calculation of the derivatives of the excess chemical potential with respect to σ and ε.
4) Prediction of the changes in the Lennard-Jones parameters as

$$\left(\Delta_{\mathrm{hyd}} G^{0}_{\mathrm{exp},i} - \Delta_{\mathrm{hyd}} G^{0}_{\mathrm{calc},i}\right) / \frac{\partial \mu^{\mathrm{PSE}\text{-}n}}{\partial \theta}. \tag{190}$$

5) Minimization of the deviation between calculation and experiment via

$$\{\Delta\sigma, \Delta\varepsilon\} = \arg\min\left[\sum_i \left(\left(\Delta_{hyd} G^{0}_{\mathrm{exp},i} - \Delta_{hyd} G^{0}_{\mathrm{calc},i}\right) - \left(\frac{\partial \mu_i^{\mathrm{PSE}\text{-}n}}{\partial \sigma}\cdot\Delta\sigma + \frac{\partial \mu_i^{\mathrm{PSE}\text{-}n}}{\partial \varepsilon}\cdot\Delta\varepsilon\right)\right)^2\right], \tag{191}$$

to obtain the change in Lennard-Jones parameters, $\Delta\sigma$ and $\Delta\varepsilon$, with the maximum and minimum of the predicted change in the respective LJ parameter as constraints and not allowing negative parameters. This Taylor-ansatz is needed since the derivatives are incomplete, the derivatives of the PMV and the electrostatic potential (from the QC calculation) are missing.
6) Change of the Lennard-Jones parameters by adding $\Delta\sigma$ and $\Delta\varepsilon$, and recalculation of the Gibbs free energies of hydration using EC-RISM.
7) Repetition of 2)-6) until convergence.

Within this procedure, the dependence of the partial molar volume on the σ and the influence of the new Lennard-Jones parameters on the solvent distribution and the partial charges are considered, but the PMV-correction parametrized using the original GAFF is used.

3.2.3.2 Calculation of tautomer populations using NMR chemical shifts

Experimental NMR chemical shifts are mostly ensemble averages over all accessible underlying conformational, tautomer and ionization states. Besides, often ^{1}H chemical shifts are not available for the protons involved in a tautomerization process because of the fast exchange between solute and solvent protons in aqueous media. Nevertheless, these experimental shifts can be used to extract the underlying tautomer fractions when supported by calculations. Therefore, the spectra of these underlying conformational, tautomer or ionization state fractions have to be calculated. To calculate the NMR chemical shift of a single conformation, the shielding constant of the solute and a suitable reference substance have to be calculated as described in *2.2.5*. The choice of a suitable reference substance and different referencing methods are discussed in chapter 4.4. The chemical shift is computed by subtracting the shielding constant of the solute from the reference shielding:

$$\delta_i = \sigma_{\text{ref}} - \sigma_i \,. \tag{192}$$

The reference substance according to the IUPAC is tetramethylsilane (TMS) in deuterated chloroform which can be used as a reference for each nucleus using the suitable Ξ-factor. In water, sodium trimethylsilylpropanesulfonate (DSS), a more water soluble TMS derivative, is the reference substance. To overcome inconsistencies from the use of different solvent models for the calculation of solute and reference substance, and therefore to ensure error cancelation by the use of the same level of theory, the secondary standard approach was used[89]

$$\delta_i = \sigma_{\text{sec,ref}} - \sigma_i + \delta_{\text{sec,ref}} \,. \tag{193}$$

Here, the experimental chemical shift difference $\delta_{\text{sec,ref}}$ between the secondary standard $\sigma_{\text{sec,ref}}$ and TMS in deuterated chloroform, respectively liquid ammonia for ^{15}N chemical shifts, is added to the calculation of chemical shifts. An NMR calculation provides the shielding constants for the respective conformation. To calculate the NMR spectrum of a tautomer, the populations of the conformations within the particular tautomer state have to be known. These populations can be calculated with the energetics, also provided by the calculation, as described in chapter 2.6. The shielding constants of the tautomer are[277]

$$\sigma_{it|i} = \sum_{c} x_{itc|it} \sigma_{itc|it} \,. \tag{194}$$

The shielding constants here are always the arithmetic means of topologically identical nuclei, which are not distinguishable from an NMR experiment. In the same fashion, the shielding constants of a specific ionization state can be calculated using the shielding constants and populations within this ionization state

$$\sigma_i = \sum_{t} \sum_{c} x_{itc|i} \zeta_{itc|i} \,. \tag{195}$$

The chemical shifts resulting from these shielding constants can directly be used for comparison with experiments measured at the respective ionization state. The shifts calculated from the tautomer shielding constants can now be used for the extraction of tautomer populations from the experimental shifts of the respective ionization state employing the multilinear regression model

$$\delta_{i,\exp} = \sum_{t} x_{it|i,\exp} \delta_{it|i,\text{calc}} \,, \tag{196}$$

with two constraints: The sum of the coefficients x has to be 1 and no negative values are allowed for them. It is important that the system of equations is not underdetermined in order to obtain a solution, therefore at least the same number of chemical shifts and tautomers have to be taken into account. The chemical shifts of the different types of nuclei show large variations w.r.t their absolute values, therefore a normalization is used to avoid the domination of the fit by a single chemical shift with a large absolute value

$$\delta_i^{\mathrm{norm}} = \delta_i / \left(\frac{1}{n} \sum_{i=1}^{n} \delta_i \right), \tag{197}$$

where n is the number of chemical shifts of the respective type of nuclei. Another possible issue is the dominance of more frequently occurring types of nuclei within the fitting process. To overcome this, another type of normalization is introduced by also taking into account the number of shifts of the respective type of nuclei via

$$\delta_i^{\mathrm{Norm.}} = \delta_i / \left(\left(\frac{1}{n} \sum_{i=1}^{n} \delta_i \right) / n \right). \tag{198}$$

A similar workflow was used in the master's thesis of the author[277] to determine the tautomer fractions of the neurotransmitter histamine. This benchmark system was also employed for this work, but on a computationally more reliable basis. The level of theory and pK_a-model developed during the SAMPL6 challenge were applied here, and especially the calculation of NMR parameters at the MP2 level of theory should improve the results.[278] The focus in this work was not only the determination of the tautomeric preferences of the histamine itself; rather, it is systematically investigated which of the nuclei are best suited for the use within the combined computational and experimental workflow for tautomer predictions.

3.2.4 Calibrating the prediction of NMR chemical shifts

The trimethylamine N-oxide (TMAO), ammonia and N-methyl-acetamide (NMA) conformations (NMA has two conformations, *cis*- and *trans*-NMA) were treated using the workflow described in chapter 3.1 with Gaussian 16.

The DSS conformations were generated by first performing an exhaustive dihedral scan at the B3LYP/6-311+G(d,p) level of theory using Gaussian 16, followed by a PCM geometry optimization as described in 3.1. All structures within a Cartesian RMSD of less than 0.8 Å were condensed into one minimum conformer, resulting in 7 different DSS minimum conformers, which were used for EC-RISM calculations. For the TMAO calculations, the Lennard-Jones parameters were taken from Ref. 279. All structures, FF parameters, energies and shielding constants for the systems are given in the SI part 04.

3.2.5 Tautomerism of nucleobases: Natural nucleobases, Hachimoji bases and derivatives

The structures of the tautomers were generated using Avogadro 1.2,[186,280] all rotamers were generated manually. The calculations were done using the computational workflow described in chapter 3.1 using Gaussian 09.[154] NMR calculations were not performed. Besides ambient conditions, pressure (1, 100, 500, 1000, 2000, 3000, 4000, 5000, 7500 and 10000 bar at 298.15 K) and temperature (278.15, 283.15, 288.15, 293.15, 298.15, 303.15, 308.15, 313.15, 318.15, 323.15, 328.15, 333.15, 338.15, 343.15, 348.15 and 363.15 K at 1 bar) dependent EC-RISM calculations were done.

3.2.6 Nucleotides at high hydrostatic pressure: NMR of adenosine monophosphate

The structures of adenosine monophosphate (AMP) were taken from the protein database (PDB, https://www.rcsb. org/ligand/AMP) in the *anti*-S-*gt* conformation. The torsion angles were manually rotated to the *anti*-S-*gg* conformation with angles O4′-C1′-N9-C4=-150° and O-C5′-C4′-O1′=-65°. The ribose conformation in the N-state was modified manually from the ribose in the GDP structure in the PDB (https://www. rcsb.org/ligand/GDP) to obtain the *anti*-N-*gg* conformer with the same torsion angles as the *anti*-S-*gt* conformation. The geometry optimizations (using Gaussian 16 rev B01) and high pressure EC-RISM protocols were chosen according to chapter 3.1. The structures, energetics and NMR parameters for each conformation and pressure are given in the SI part 06.

3.2.7 Nucleotides at various temperatures

In contrast to nucleobases, which are very rigid due to the aromatic ring system, nucleotides have many degrees of freedom. The base and the phosphate chain are bound to the (desoxy-)ribose by a rotatable bond allowing a lot of orientations, and the phosphate chain itself is very flexible. Additionally, the ribose has multiple conformations of the ring and the hydroxyl groups. Therefore, a proper conformational sampling is necessary for the calculation of nucleotide energetics and NMR parameters. In this work, a combined workflow using quantum chemical calculations and molecular dynamics simulations with an implicit solvent model was applied.

At first, the molecular structure of the respective tautomer of the nucleotide was generated with the molecular editor Avogadro 1.2[186,280] and optimized utilizing molecular mechanics applying the steepest descent algorithm and the MMFF94s[183] FF as implemented in Avogadro. This structure was submitted to a quantum chemical geometry optimization at the HF/6-31+G(d) level of theory in Gaussian 16 with tight convergence criteria, followed by a single point calculation at the same level of theory in which the internal option for the detailed output of the electrostatic potential has to be activated in Gaussian 16 (by using the flag IOp(6/33=2) in the head line). This is needed for the calculation of atomic partial charges with the restrained electrostatic potential (RESP) model.[281,282,283] This model has some advantages compared to partial charges based on semi-empirical models like AM1-BCC[75,268] by an accurate description of the buried atoms far from the molecular surface, which are often badly represented by the fitting process because the electrostatic potential is dominated by the surface atoms, which can lead to large variations of the partial charges of the buried atoms. The RESP charges are well suited for the use of aqueous simulations due to the overpolarization of the molecule using the HF/6-31+G(d) level of theory which represents the situation in aqueous solution quite well. Therefore, the espgen and respgen utilities of the antechamber[284] program package implemented in Amber 18[190] were used. Afterwards, antechamber was used for the parametrization of the molecule with the GAFF[184] FF version 1.7 and the data was submitted to a molecular dynamics simulation with the sander program of Amber 18. The simulation was performed for 100 ns with a 1 fs timestep at 500 K using Langevin dynamics and the ALPB solvation model. Every 10000th snapshot was stored for subsequent geometry optimization to the next-nearest local minimum with a convergence criterion of 10^{-4} kcal mol^{-1} Å^{-1} for the maximum gradient norm with Amber18. All of the minimized structures within a Cartesian RMSD of 0.025-times the number of solute atoms were condensed to the minimum structure of the cluster and used for quantum chemical geometry optimizations at the B3LYP/6-311+G(d,p)/PCM level of theory in Gaussian 16, equivalent to the SAMPL6 settings. These structures were again condensed like described above, and the 10 minimum structures were used for the EC-RISM calculations. This workflow is based on the one used by P. Kibies for the generation of conformations for the WZ4002[82] but a modernized and more flexible version.

The structures of all nucleosides, mono- and triphosphate nucleotides were generated for each of the base tautomers given in chapter 4.5; thereby, the phosphate chains are deprotonated, resulting in ionization states of 0 for the nucleosides, -2 for the monophosphates and -4 for the triphosphates. For each of these species (in total 45, by having 5 nucleobases with each 3 tautomers as nucleoside, mono- and triphosphate) the workflow described above was performed. This resulted in a very large number of quantum chemical calculations needed, making the investigation of the influence of the ionization state, by using the workflow on nucleotides at different ionization states, unfeasible in this work due to limited computational resources. The number of resulting conformations after the PCM optimization for each tautomer is given in Table 3. The input structures were all chosen with the 'standard' stereochemistry and conformation. These conformations were used for temperature dependent EC-RISM calculations at 278.15, 283.15, 288.15, 293.15, 298.15, 303.15, 308.15, 313.15, 318.15, 323.15, 328.15, 333.15, 338.15, 343.15, 348.15, 363.15, 368.15 and 372.756 K, making it computationally costly to use all conformations for the EC-RISM calculations. To account for the conformational flexibility of the molecules, while simultaneously reducing the computational effort, the 10 energetically lowest structures from the PCM optimization were used for EC-RISM calculations, resulting in 180 calculations per species. The calculation of NMR parameters is computationally very demanding at the MP2 level of theory since no frozen core approximation can be used, and especially the phosphorylated species have a high number of electrons. This results in swap files of multiple terabytes for a single calculation on a nucleoside monophosphate conformation when using 0.5 TB memory in Gaussian 16, so calculations on each minimum used for EC-RISM are computationally impossible. Thus, the populations of the conformations of each species at each temperature were calculated, and only the species contributing for up to 99 % of the conformational ensemble of the respective molecule at the respective temperature were submitted to NMR calculations with a maximum of 3 conformations per molecule and temperature. The calculation of nucleoside triphosphate NMR parameters was not possible using the given computational resources. Given the size of the molecules and the fact that all of the tautomerizing protons are in six-membered rings, for which the calculations performed well on the SAMPL2 dataset even without the inclusion of additional high-level gas-phase energies in the thermodynamic cycle, coupled cluster calculations were not done for the nucleosides and nucleotides. Additionally, the inclusion of coupled-cluster energies does not affect the tautomer populations of histamine, but only the distribution in the underlying conformational ensemble is affected, which makes coupled-cluster calculations unnecessary for the tautomer elucidation of nucleosides and nucleotides.

Table 3: Number of PCM minima of the investigated species, the nomenclature of the tautomeric states of the nucleobases in the respective nucleoside or nucleotide species is according to chapter 4.5. The conformations were generated using the workflow described in chapter 3.2.7 and are used for EC-RISM/NMR calculations.

Tautomer	Nucleoside	Nucleoside monophosphate	Nucleoside triphosphate
A	22	53	217
N1-A	23	55	247
N3-A	9	21	121
G	25	65	194
Enol-G	23	81	117
N3-G	32	50	353
U	44	37	279
2-Enol-U	92	63	100
4-Enol-U	69	66	178
T	22	72	57
2-Enol-T	16	21	100
4-Enol-T	32	110	158
C	27	56	236
Enol-C	43	57	102
N3-C	23	45	200

Since the populations of all tautomers and conformations were calculated via EC-RISM, the results include the effect of the different environmental conditions. All of the raw data, including molecular structures, EC-RISM results and NMR parameters are given in the SI part 07.

Experimental details

The NMR experiments were all performed by the author during a RESOLV internship in Roland Sigel's lab at the University of Zürich in Switzerland. ^{13}C and ^{15}N labeled nucleosides (adenosine, guanosine, thymidine and cytidine, all as desoxyribonucleosides) and nucleotides (ATP, GTP, UTP and CTP, all as ribonucleotides) were purchased from Silantes and used without further purification. The measurements were done using a 700 MHz Bruker Avance III Neo spectrometer with Cryo TXI probe. Direct measurements of the ^{13}C, ^{15}N, and ^{1}H (using water suppression) nuclei resonances as well as ^{1}H/^{13}C-HSQC experiments of the sugar region (to support the assignment of the ^{1}H resonances in the sugar region, which is difficult due to interference with the water signal) were performed in water/D_2O (80:20) at 278.15, 298.15, 323.15 and 348.15 K.

The temperature calibration was done according to the manufacturer's specifications using methanol and ethylene glycol yielding the spectral ratios given in Table 4. These spectral ratios are needed to account for the temperature dependence of the reference frequency. An internal reference was used, which was calibrated at 298.15 K, the spectral ratios modify the reference frequency to yield the correct temperature dependent reference.

Table 4: Spectral ratios (in Hz) used for the referencing of temperature dependent chemical shifts. Determined by measuring methanol and ethylene glycol on the same 700 MHz Bruker Avance III Neo spectrometer as used for the nucleoside and nucleotide measurements.

Temperature	^{1}H	^{13}C	^{15}N
278.15 K	-209.47	-521.74	-22.65
298.15 K	-49.78	-481.58	-6.47
323.15 K	149.84	-431.39	13.76
348.15 K	349.46	-381.20	33.99

IV. Results

4.1 EC-RISM for temperature variations: Parametrization and benchmarking

EC-RISM with the PMV correction was successfully applied to calculations in water[4,5] and different solvents like cyclohexane[232] and dry and water saturated octanol;[8] it is even possible to use it for high pressure calculations in water.[7] To further increase the applicability of EC-RISM and give rise to the possibility to study the tautomerism of nucleic acid building blocks under a broader variety of environmental and especially early life conditions, the parametrization of a PMV correction for temperature variations is crucial. In the final chapter of this work, the tautomerism of nucleosides and nucleotides is investigated computational and using temperature dependent NMR experiments. Therefore, it is necessary that such an EC-RISM parametrization covers the whole range of temperatures at which the experiments are performed. The results of the parametrization are shown in Table 5 and Figure 7; the full data, also splitted into the separate components, are given in the Energies subfolder of SI part 01.

Table 5: Statistical metrics for Gibbs free energies of solvation calculated by EC-RISM and by TI in comparison with experimental values for 298.15 K and for Gibbs free energies of solvation calculated by EC-RISM with and without additional temperature correction in comparison with the temperature dependent reference data (root mean square error RMSE, mean absolute error MAE, mean signed error MSE, slope m', intercept b', and coefficient of determination R^2 from descriptive regression) of the training set.

Model	RMSE	MAE	MSE	*m*'	*b*’	R^2
TI(298.15K)	1.56	1.24	0.65	0.87	-1.28	0.94
EC-RISM^0(298.15K)	1.98	1.32	-0.86	0.88	0.12	0.89
EC-RISM^0(278.15K)	1.34	1.23	-1.23	1.02	1.09	0.99
EC-RISM^T(278.15K)	0.77	0.70	-0.70	1.01	-0.60	1.00
EC-RISM^0(283.15K)	0.98	0.89	-0.89	1.02	-0.75	0.99
EC-RISM^T(283.15K)	0.56	0.49	-0.49	1.01	-0.38	1.00
EC-RISM^0(288.15K)	0.63	0.58	-0.57	1.01	-0.50	1.00
EC-RISM^T(288.15K)	0.36	0.31	-0.31	1.01	-0.25	1.00
EC-RISM^0(293.15K)	0.31	0.28	-0.28	1.01	-0.22	1.00
EC-RISM^T(293.15K)	0.18	0.15	-0.14	1.01	-0.10	1.00
EC-RISM^0(303.15K)	0.26	0.24	0.24	1.00	0.24	1.00
EC-RISM^T(303.15K)	0.14	0.12	0.10	1.00	0.12	1.00
EC-RISM^0(308.15K)	0.50	0.45	0.45	0.99	0.41	1.00
EC-RISM^T(308.15K)	0.24	0.20	0.18	1.00	0.16	1.00

Model	RMSE	MAE	MSE	*m*'	*b*’	R^2
EC-RISM^0(313.15K)	0.74	0.67	0.67	0.99	0.59	1.00
EC-RISM^T(313.15K)	0.33	0.27	0.27	0.99	0.22	1.00
EC-RISM^0(318.15K)	0.89	0.83	0.83	0.99	0.75	1.00
EC-RISM^T(318.15K)	0.33	0.29	0.29	0.99	0.25	1.00
EC-RISM^0(323.15K)	1.06	0.98	0.98	0.98	0.88	0.99
EC-RISM^T(323.15K)	0.35	0.31	0.30	1.00	0.25	1.00
EC-RISM^0(328.15K)	1.22	1.14	1.14	0.98	1.05	0.99
EC-RISM^T(328.15K)	0.37	0.33	0.33	0.99	0.29	1.00
EC-RISM^0(333.15K)	1.34	1.24	1.24	0.98	1.12	0.99
EC-RISM^T(333.15K)	0.35	0.29	0.29	0.99	0.23	1.00
EC-RISM^0(338.15K)	1.45	1.35	1.35	0.98	1.24	0.99
EC-RISM^T(338.15K)	0.31	0.28	0.27	0.99	0.21	1.00
EC-RISM^0(343.15K)	1.52	1.43	1.43	0.98	1.35	0.99
EC-RISM^T(343.15K)	0.25	0.21	0.21	1.00	0.19	1.00
EC-RISM^0(348.15K)	1.62	1.52	1.52	0.98	1.41	0.99
EC-RISM^T(348.15K)	0.23	0.18	0.16	0.99	0.12	1.00
EC-RISM^0(363.15K)	1.76	1.64	1.64	0.98	1.54	0.98
EC-RISM^T(363.15K)	0.22	0.17	-0.13	1.00	-0.15	1.00

Compared to the experimental values from the MNSOL at 298.15 K, the TI and EC-RISM yield an RMSE of 1.56 and 1.98 kcal/mol, respectively. Even though the RMSE of the TI calculations is smaller, the error of both methods is in a comparable range, underlining the validity of the chosen approach using temperature dependent TI calculations as reference data for the parametrization. As expected from the fitting process, the EC-RISM^T results are in better agreement with the reference data than the uncorrected EC-RISM calculations. They result in smaller errors (RMSE, MAE and MSE) for all temperatures (except 298.15 K where the results are the same by definition), show smaller intercepts and slopes, and coefficients of determination closer to 1. This indicates that the fitting was successful. The results of the lowest and highest temperature used for the parametrization (278.15 K and 363.15 K), and the errors for all temperatures are depicted in Figure 7.

Table 6: Statistical metrics for Gibbs free energies of solvation calculated by EC-RISM with and without additional temperature correction in comparison with the temperature dependent experimental data (root mean square error RMSE, mean absolute error MAE, mean signed error MSE, slope m', intercept b', and coefficient of determination R^2 from descriptive regression) of the test set.

Model	RMSE	MAE	MSE	*m*'	*b*'	R^2
EC-RISM^0(278.15 K)	1.82	1.56	-1.55	1.11	-1.22	0.96
EC-RISM^T(278.15 K)	1.51	1.17	-1.11	1.11	-0.78	0.95
EC-RISM^0(283.15 K)	1.60	1.29	-1.25	1.10	-0.94	0.96
EC-RISM^T(283.15 K)	1.40	1.03	-0.93	1.10	-0.62	0.95
EC-RISM^0(288.15 K)	1.43	1.07	-0.99	1.10	-0.71	0.95
EC-RISM^T(288.15 K)	1.32	0.96	--0.77	1.10	-0.49	0.95
EC-RISM^0(293.15 K)	1.31	0.95	-0.76	1.10	-0.50	0.94
EC-RISM^T(293.15 K)	1.27	0.94	-0.65	1.09	-0.39	0.94
EC-RISM^0(298.15 K)	1.24	0.93	-0.55	1.09	-0.31	0.94
EC-RISM^T(298.15 K)	1.24	0.93	-0.55	1.09	-0.31	0.94
EC-RISM^0(303.15 K)	1.21	0.92	-0.36	1.08	-0.14	0.93
EC-RISM^T(303.15 K)	1.23	0.93	-0.47	1.08	-0.25	0.93
EC-RISM^0(308.15 K)	1.22	0.93	-0.20	1.08	0.00	0.93
EC-RISM^T(308.15 K)	1.22	0.92	-0.37	1.08	-0.22	0.93
EC-RISM^0(313.15 K)	1.24	0.97	-0.06	1.08	0.12	0.92
EC-RISM^T(313.15 K)	1.21	0.92	-0.39	1.08	-0.20	0.93
EC-RISM^0(318.15 K)	1.28	1.01	0.06	1.07	0.23	0.91
EC-RISM^T(318.15 K)	1.21	0.93	-0.37	1.07	-0.21	0.93
EC-RISM^0(323.15 K)	1.32	1.07	0.17	1.07	0.32	0.91
EC-RISM^T(323.15 K)	1.22	0.93	-0.38	1.07	-0.23	0.93
EC-RISM^0(328.15 K)	1.36	1.12	0.25	1.06	0.39	0.90
EC-RISM^T(328.15 K)	1.22	0.93	-0.40	1.06	-0.27	0.93
EC-RISM^0(333.15 K)	1.40	1.17	0.32	1.06	0.44	0.90
EC-RISM^T(333.15 K)	1.23	0.94	-0.45	1.06	-0.33	0.93
EC-RISM^0(338.15 K)	1.44	1.20	0.37	1.06	0.48	0.89
EC-RISM^T(338.15 K)	1.24	0.95	-0.50	1.05	-0.40	0.93
EC-RISM^0(343.15 K)	1.46	1.23	0.41	1.05	0.50	0.89
EC-RISM^T(343.15 K)	1.26	0.97	-0.58	1.05	-0.49	0.93
EC-RISM^0(348.15 K)	1.48	1.25	0.43	1.05	0.51	0.88
EC-RISM^T(348.15 K)	1.29	1.00	-0.67	1.04	-0.60	0.93
EC-RISM^0(363.15 K)	1.49	1.25	0.40	1.04	0.45	0.88
EC-RISM^T(363.15 K)	1.46	1.15	-1.03	1.03	-1.00	0.93

To evaluate the predictive power of EC-RISM^T, an independent test set is needed. The test set is built as a subset of the Chamberlin dataset. As described above, the experimental data in this dataset are not given for different temperatures, only the highest and lowest experimental temperature available and the corresponding values are given for each molecule. To correlate the calculated data for each data point with experimental values, an approximation is needed, therefore a linear approximation of the Gibbs free energies of hydration was applied using the given points (the minimum and maximum temperatures, the respective Gibbs free energies of hydration as well as the linear regression parameters are given in the appendix in Table 48). Such an approximation is neglecting the temperature dependence of Δ*H* and Δ*S* within the Gibbs free energy and is, in reference to Chamberlin et al.,[252] called van't Hoff model. This way, the slope of the linear regression is the negative of the solvation entropy, which could be used to validate the approach, but unfortunately is not given in the literature. The comparison between the reference data from this van't Hoff model and the computed results using EC-RISM and EC-RISM^{T} are given in Table 6 and Figure 7.

RMSE and MAE of the test set are comparable between EC-RISM and EC-RISM^{T} for temperatures near to 298.15 K and, unexpected from the results for the training set, at the highest temperature, 363.15 K. For all other temperatures, EC-RISM^{T} outperforms EC-RISM by showing a small RMSE and MSE over a broad range of temperatures. Despite the overall better performance of EC-RISM^{T}, the native EC-RISM covers the temperature dependence quite well, especially compared to the performance at high pressures. The MSE of the EC-RISM^{T} results has a parabolic form and is always negative while, in contrast, the one of the EC-RISM results has a zero crossing at approximately 320 K and an overall steadier trend, which is likely caused by a monotonous trend of the temperature dependent PMVs which is not compensated by the additional temperature dependent parameter. This results in small errors for the test set at 363 K and is not in agreement with the results of the training set, for which the MSE of the EC-RISM results is worse than for the result calculated with EC-RISM^{T}. This may be caused by the choice of molecules for the test set: they are very small (the largest compound contains 11 non-hydrogen atoms) because no conformational sampling is done for the test set, so the temperature dependent parameter may be too large for this test set since it is parametrized from a training set of small, medium sized and large (up to 25 heavy atoms) molecules to be applicable to a broad range of small molecules. A possible explanation for the training set showing slightly higher errors compared to the training set performance is the neglect of conformational and tautomeric freedom of the test set molecules as well as not considering possible ionization states which may play an important role at ambient conditions.

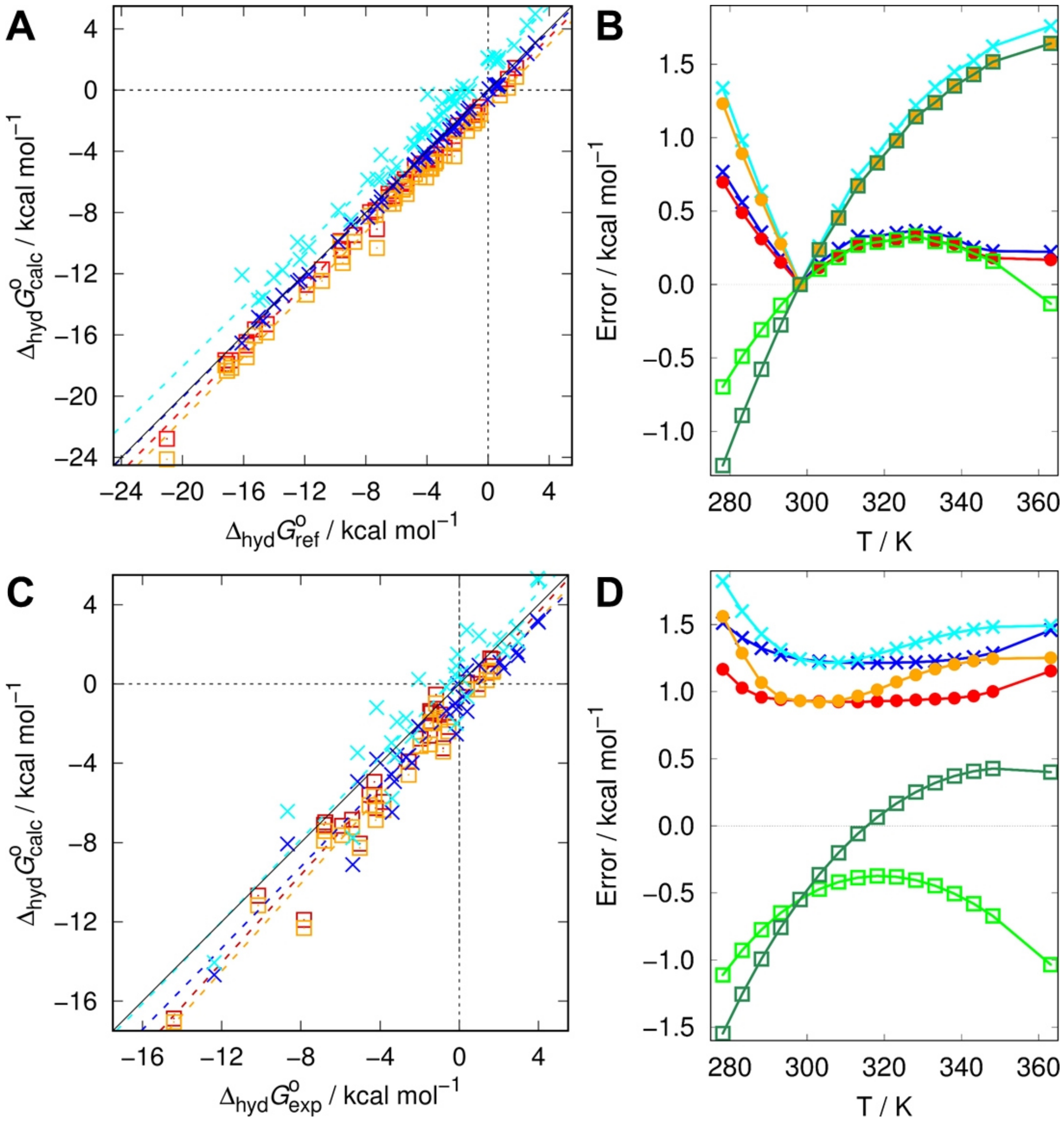


Figure 7: Results of the calculations of the training (A and B) and test set (C and D) used for the parametrization and validation of the temperature dependent PMV correction for EC-RISM. On the left side (A and C), the calculated vs. reference data (at 278.15 K (crosses) and 363.15 K (squares) with EC-RISM^T in blue and red, and EC-RISM in cyan and orange, on the right side (B and D) the trends of the errors (RMSE (crosses), MAE (points) and MSE(squares) EC-RISM^T in blue, red, and green, EC-RISM in cyan, orange and dark-green) are shown.

Overall, the newly introduced temperature dependent PMV correction for EC-RISM slightly outperforms the correction for ambient conditions for the training and the test set, and it is recommended to use this correction for temperature dependent calculations of small molecules in liquid water. Although the MSE implies that EC-RISM might work better at higher temperatures, the linear regression parameters indicate that EC-RISM^T is the more stable approach. A major part of this work is the calculation of the nucleobase, nucleoside and nucleotide tautomerism (Chapters 4.5 and 4.7) over a range of temperatures, which is now possible with this correction. For these investigations, temperature dependent NMR spectra of nucleosides and nucleotides were recorded in a temperature range covered by the new correction, making the correlation between calculated and experimental results possible.

4.2 Calibrating tautomer predictions: the SAMPL2 dataset

Before calculating the tautomer stability of nucleic acid building blocks, especially at extreme conditions, it is needed to benchmark the performance of EC-RISM for tautomer predictions at ambient conditions. Therefore, the SAMPL2 dataset is used, this dataset is introduced in chapter 1.4. The SAMPL2 challenge[90] was part of the "statistical assessment of the modeling of proteins and ligands" blind prediction series back in 2009. During the original challenge, the Kast group participated, resulting in an RMSE of 0.57 kcal/mol for the explanatory set, containing the 5-ring compounds, and 2.91 kcal/mol for the 6-ring compounds of the obscure dataset, resulting in an overall RMSE of 2.0 kcal/mol.

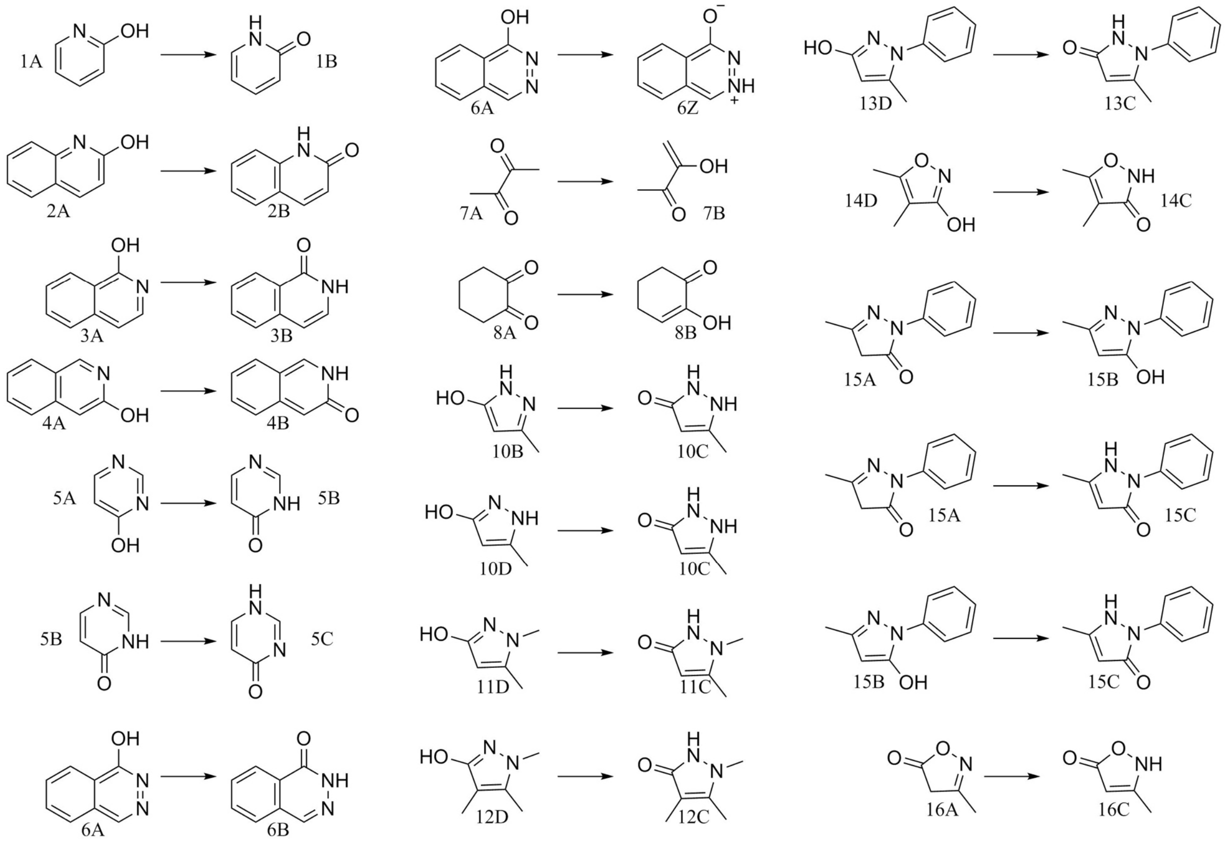


Figure 8: Compounds of the SAMPL2 dataset. All pairs of the obscure (1-6) and explanatory (8-16) datasets for which experimental energy differences are available are given. It is important to notice that compounds 7 and 8, which are part of the explanatory set but no 5-ring compounds, are only considered in the overall statistics.

At that time, the EC-RISM solvation model could be used to calculate energy differences with high accuracy; however, absolute values like solvation free energies could not be calculated reliably due to the lack of a PMV correction term. Therefore, the Kast group did not participate in the challenges SAMPL3[285] and SAMPL4,[286] where the blind prediction of hydration free energies was the task. The workgroup participated again in the SAMPL challenges 5,[232,287] 6,[4,288] 6.2[8,289] and 7 because the PMV correction, developed by D. Tomazic et al., gave access to the calculation of absolute values. The participation in the SAMPL challenges was always in close collaboration with co-workers from Sanofi, especially S. Güssregen and is reviewed in Ref. 10. Using the computational workflow established during the post submission phase of the SAMPL6 challenge, described in chapters *2.5.2* and 3.1 as well as Ref. 4, aqueous pK_a values can be calculated with an RMSE of 1.04 pK_a units, employing the pK_a model described in chapter 2.6 and using the PMV correction, an RMSE of 2.04 kcal/mol for the MNSOL database (1.56 kcal/mol for neutral molecules, all molecules in the SAMPL2 dataset are neutrals) could be obtained. This work is mainly concerned with tautomerism, and the SAMPL2 dataset is one, if not the most well-known dataset for tautomer predictions, so that the new computational workflow from the SAMPL6 challenge had to be tested on this dataset. For this task, the thermodynamic cycle for the calculation of tautomer transitions had to be considered that has been developed in our analysis of the tautomer stability of the Hachimoji code in Ref. 9.

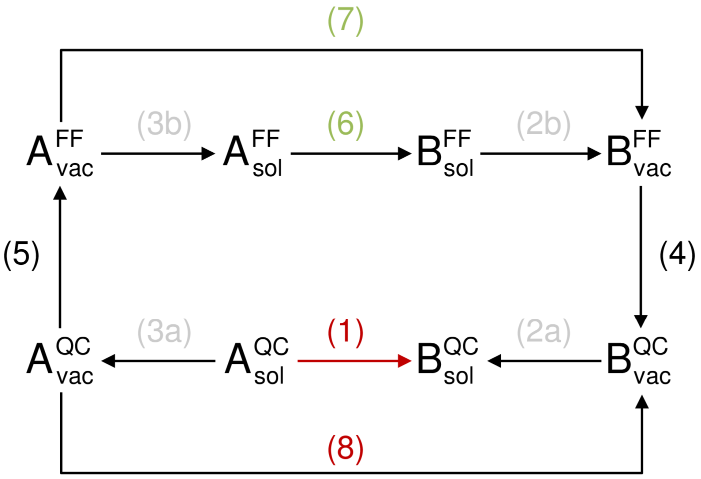


Figure 9: Thermodynamic cycle for the calculation of the tautomer ratio between tautomers A and B in a mixed FF / quantum-chemical framework for determining the free energy difference between the tautomers in solution (subscript sol) and the gas phase (subscript vac). The desired path (1) for the calculation of the quantum chemical tautomer ratio in solution can be obtained by calculation of different paths, they are in detail described in the main text. Developed by Stefan M. Kast and published in Ref. 9 (https://pubs.acs.org/doi/abs/10.1021/acs.jctc.9b01079, the article on request link is given in 290.)

Using the thermodynamic cycle in Figure 9 and assuming the same standard states for the molecules, like infinite dilution at a specified temperature and pressure, there are multiple ways to calculate $\Delta G(1)$, described by the red arrow. The direct approach is to calculate the absolute free energies of both species A and B in solution using a quantum chemical method, such as EC-RISM, and calculate the difference:[9]

$$\Delta G^{(1)} = \Delta G\left(\mathrm{B}_{\mathrm{sol}}^{\mathrm{QC}}\right) - \Delta G\left(\mathrm{A}_{\mathrm{sol}}^{\mathrm{QC}}\right). \tag{199}$$

This direct approach can be used together with computational methods providing accurate absolute free energies in solution. If the method is able to compute accurate hydration free energies, for example by being parametrized with respect to experimental ones, but does not cover the absolute intramolecular energy well by being parametrized for low level QC theories like HF, which is often the case for continuum models like PCM, an indirect method can be used. This can be done by calculating the hydration free energies for both tautomers, ΔG(3a) and ΔG(2a), using the low-level computational method and combining it with the gas-phase free energy difference from a high-level calculation like coupled cluster according to

$$\Delta G^{(1)} = \Delta G_{\text{low level}}^{(3\mathrm{a})} + \Delta G_{\text{low level}}^{(2\mathrm{a})} + \Delta G_{\text{high level}}^{(8)}. \tag{200}$$

It is also possible to use the indirect approach together with FF-based methods, such as TI calculations from molecular dynamics simulations. Consequently, the thermodynamic cycle, which incorporates two different Hamiltonians, the quantum chemical and the FF, needs to be closed. This is achieved by assuming the equivalence of both Hamiltonians to reproduce the hydration free energies of both tautomers

$$\Delta G^{(2\mathrm{a})} + \Delta G^{(2\mathrm{b})} = \Delta G^{(3\mathrm{a})} + \Delta G^{(3\mathrm{b})} = 0. \tag{201}$$

Now, the hydration free energy can be calculated using any FF-based method. In this work, TI calculations using molecular mechanics simulations are employed, but Monte-Carlo sampling can also be used[9]

$$\Delta G^{(1)} = -\Delta G_{\mathrm{TI}}^{(3\mathrm{b})} - \Delta G_{\mathrm{TI}}^{(2\mathrm{b})} + \Delta G_{\text{QC,high level}}^{(8)}. \tag{202}$$

This approach is the so-called dual topology approach, here the hydration free energies of both tautomers have to be computed with a respective TI calculation. It is also possible to calculate the hydration free energy difference between both tautomers within a single MD/MC calculation using alchemical transformations to switch from one to the other tautomer during the calculation. The free energy difference in solution is given by

$$\Delta G^{(1)} = \Delta G^{(6)} + \Delta G^{(5)} + \Delta G^{(4)} = \Delta G_{\mathrm{TI}}^{(6)} - \Delta G_{\mathrm{TI}}^{(7)} + \Delta G_{\text{QC,high level}}^{(8)}. \tag{203}$$

The calculations for the SAMPL2 dataset were done using the direct approach and the indirect together with vacuum coupled cluster calculations at the CCSD(T)/cc-pVTZ level of theory and using thermal corrections.

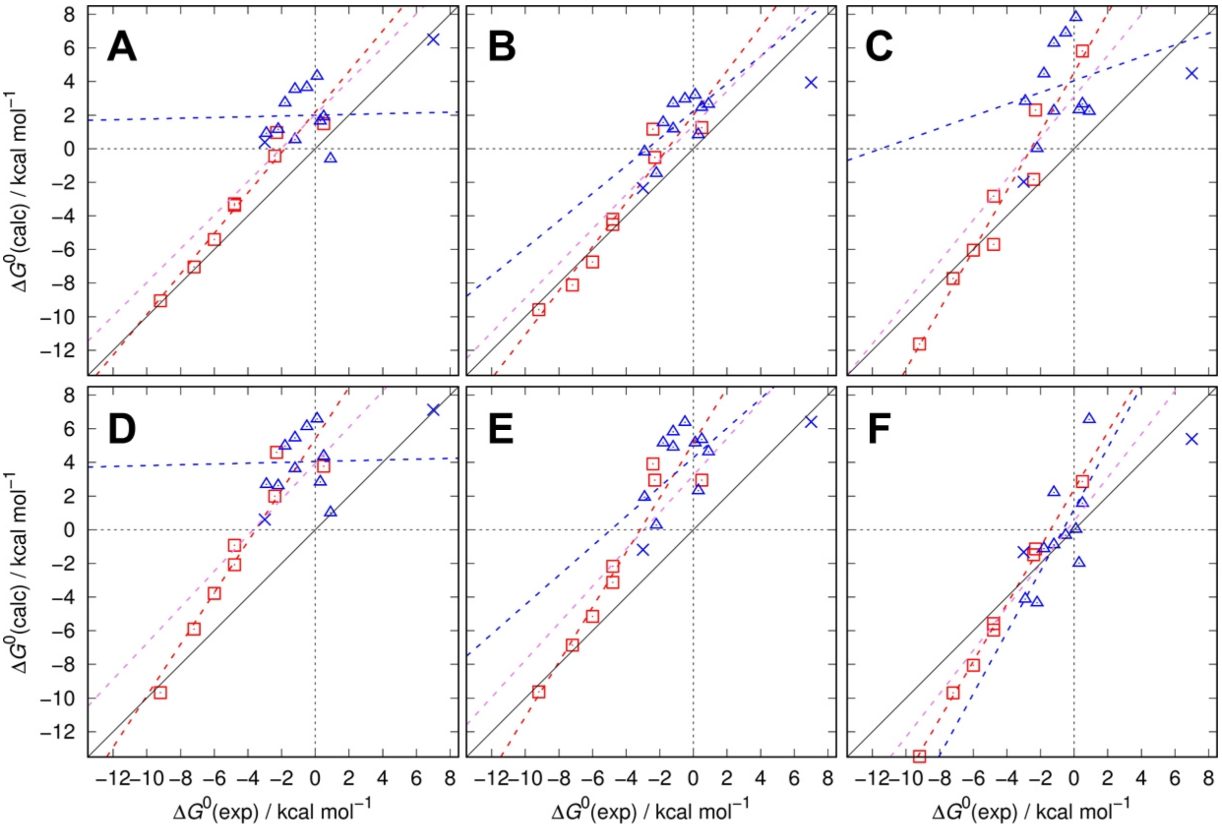


Figure 10: Results of the calculations on the SAMPL2 dataset. The direct approach is shown in A (MP2/6-311+G(d,p)/EC-RISM), D (MP2/6-311+G(d,p)/PCM), and F (B3LYP/6-311+G(d,p)/PCM), the indirect in B (MP2/6-311+G(d,p)/EC-RISM+CCSD(T)), C (TI+CCSD(T)), and E (MP2/6-311+G(d,p)/PCM+CCSD(T)); CCSD(T) implicitly denotes the inclusion of thermal corrections at the level of theory of the structure optimization (B3LYP/6-311+G(d,p)). Blue triangles depict the 5-membered rings of the explanatory dataset, red squares the 6-membered rings (obscure dataset), the dashed lines show the corresponding linear regressions; the blue crosses are the explanatory pairs 7 and 8 which are included in the overall regression (violet line). Statistical parameters can be found in Table 7.[10]

The results of the calculations on the SAMPL2 dataset are shown in Table 7 and Table 8, and are depicted in Figure 10. Multiple levels of theory were used to investigate this dataset: EC-RISM using the computational approach developed for the SAMPL6 challenge was tested with and without the inclusion of high-level gas-phase calculations. The same was done using PCM at the MP2 and B3LYP level of theory, and frozen geometry dual-topology TI calculations were conducted. The overall RMSE is the lowest for EC-RISM/CCSD(T), followed by PCM/B3LYP. The latter is an exceptional level of theory here because for all of the other ones, the RMSE, MAE, and MSE are lower for the 5-membered rings than for the 6-membered rings, and each of these metrics decreases, both overall and for the respective 5- and 6-membered rings, upon inclusion of the high-level gas-phase term, while for PCM/B3LYP, this inclusion worsens the results. Therefore, these results are neglected in the following section.

Especially the quality of the tautomer predictions of the 5-membered rings increases drastically by including electron correlation effects at a high-level: The errors, while still higher than for the 6-membered rings, are lower by more than 0.5 kcal/mol, and the descriptive regression parameters, which show a lack of predictive power for the 5-membered rings with zero or near zero slopes and coefficients of determination, improve significantly. Generally, the use of EC-RISM is recommended over PCM/MP2 because the errors are lower by around 1.5 kcal/mol, and the slopes and coefficients of determination are closer to one. Additionally, the use of coupled-cluster gas-phase energies and thermal corrections is advocated if computationally affordable, especially when dealing with 5-membered rings. For this purpose, however, gas-phase optimizations, frequency calculations and the costly coupled-cluster calculations have to be performed.

Table 7: Results of the SAMPL2 calculations. Statistical metrics root mean square error (RMSE), mean absolute error (MAE) and mean signed error (MSE) are given in kcal/mol. Besides, the regression parameters m and b (in kcal/mol), and the coefficient of determination R^2 from descriptive regression are shown. Compounds 7 and 8, which are part of the explanatory dataset but no 5-membered rings, are considered in the overall statistic. Additionally, the results from the original SAMPL2 submission using only minimum energies (min) or the partition function approach (Z) are presented.[91] The full TI results are presented in Table 49 in the appendix.

Model	RMSE	MAE	MSE	*m*	*b*	R^2
MP2/6-311+G(d,p)/ECRISM						
All	2.69	2.24	2.04	1.00	2.04	0.79
5-membered rings	3.36	3.08	2.78	0.02	2.00	0.00
6-membered rings	1.59	1.26	1.26	1.21	2.19	0.95
MP2/6-311+G(d,p)/ECRISM//CCSD(T)						
All	2.20	1.83	1.32	1.03	1.38	0.79
5-membered rings	2.62	2.39	2.39	0.82	2.24	0.46
6-membered rings	1.52	1.13	0.62	1.31	2.02	0.93
TI//CCSD(T)						
All	4.11	3.28	2.64	1.23	3.10	0.65
5-membered rings	5.19	4.57	4.57	0.35	4.06	0.03
6-membered rings	2.75	2.05	1.07	1.77	4.54	0.94
MP2/6-311+G(d,p)/PCM						
All	4.45	3.86	3.81	1.07	3.95	0.72
5-membered rings	5.25	4.83	4.83	0.02	4.05	0.00
6-membered rings	3.65	3.14	3.02	1.53	5.41	0.91
MP2/6-311+G(d,p)/PCM//CCSD(T)						
All	3.90	3.39	3.03	1.10	3.23	0.70

Model	RMSE	MAE	MSE	*m*	*b*	R^2
5-membered rings	4.62	4.39	2.32	0.87	4.29	0.35
6-membered rings	3.24	2.57	4.39	1.63	5.16	0.92
B3LYP/6-311+G(d,p)/PCM						
All	2.34	1.77	-0.04	1.28	0.54	0.82
5-membered rings	2.38	1.70	0.55	1.83	1.21	0.52
6-membered rings	2.19	1.90	-0.79	1.70	2.40	1.00
B3LYP/6-311+G(d,p)/PCM//CCSD(T)						
All	3.21	2.79	2.37	1.16	2.68	0.78
5-membered rings	3.91	3.70	3.70	1.22	3.87	0.58
6-membered rings	2.38	1.88	1.37	1.56	3.91	0.95
SAMPL2/min						
All	1.98	1.49	-1.00	1.18	-0.64	0.86
5-membered rings	0.58	0.46	0.12	0.83	-0.02	0.78
6-membered rings	2.90	2.67	-2.67	1.00	-2.20	0.89
SAMPL2/Z						
All	1.93	1.47	-0.94	1.16	-0.63	0.86
5-membered rings	0.66	0.52	0.21	0.84	0.09	0.74
6-membered rings	2.78	2.53	-2.53	1.10	-2.10	0.89

For tautomer predictions, EC-RISM performs better than the TI, for which the errors are much higher and near to those for PCM/MP2, but the general trends observed for EC-RISM and PCM/MP2 stay the same. Noteworthy is the big difference between 5- and 6-membered rings here: Using TI calculations, the high-level gas-phase terms have to be included, the inclusion of these data leads to a reduction of this difference for EC-RISM and PCM/MP2. In contrast, this difference is even larger than for the plain EC-RISM and PCM/MP2 approaches with approximately 2.5 kcal/mol for the TI calculations. Despite being by far the most expensive of the approaches, TI calculations are not recommended for tautomer predictions based on the SAMPL2 results. The error of the solvation free energies from the TI calculations is in the range of 0.2 kcal/mol for each of the molecules in the dataset, so no systematic shift of the results can be related to this (Table 49).

Table 8: Experimental tautomerization Gibbs energies (kcal/mol) including estimated errors,[90] calculated values from original SAMPL6 setup (EC-RISM), using solvation free energies from TI calculations as well as PCM results at the B3LYP/6-311+G(d,p) and MP2/6-311+G(d,p) levels of theory. The indirect route, using Eqn. 200, is denoted with "CCSD(T)". Additionally, the results from the original SAMPL2 submission using only minimum energies (min) or the partition function approach (Z) are presented.[10,91]

Reaction	Exp.	Error	EC-RISM	EC-RISM	TI	PCM	PCM	PCM	PCM	SAMPL2	SAMPL2
				CCSD(T)	CCSD(T)	(MP2)	(MP2)	(B3LYP)	(B3LYP)	min	Z
							CCSD(T)		CCSD(T)		
1A→1B	-4.8	0.3	-3.38	-4.52	-2.83	-0.93	-2.17	-5.57	-3.16	-7.73	-7.57
2A→2B	-6.1	0.3	-5.40	-6.74	-6.04	-3.79	-5.16	-8.06	-5.99	-9.66	-9.29
3A→3B	-7.2	0.3	-7.04	-8.12	-7.73	-5.90	-6.86	-9.69	-7.58	-11.17	-11.12
4A→4B	-2.3	0.4	0.96	-0.52	2.31	4.59	2.94	-1.14	1.78	-4.57	-4.43
5A→5B	-4.8	0.5	-3.28	-4.19	-5.70	-2.08	-3.13	-5.97	-3.97	-6.16	-5.83
5B→5C	0.5	0.2	1.50	1.26	5.81	3.77	2.95	2.87	2.19	-0.51	-0.51
6A→6B	-9.2	0.4	-9.05	-9.59	-11.63	-9.67	-9.63	-13.46	-10.08	-11.15	-11.12
6A→6Z	-2.4	0.3	-0.43	1.17	-1.82	1.99	3.91	-1.49	2.09	-6.72	-6.69
7A→7B	7.0	1.5	6.50	3.94	4.49	7.12	6.41	5.39	6.73	5.11	4.71
8A→8B	-3.0	3.0	0.38	-2.34	-1.97	0.61	-1.18	-1.33	-0.28	-1.01	-1.38
10B→10C	-2.9	0.4	0.91	-0.20	2.83	2.69	1.93	-4.13	0.60	-2.84	-2.83
10D→10C	-1.2	0.2	3.54	2.70	6.29	5.45	5.81	-0.89	4.50	-0.55	-0.45
11D→11C	-0.5	0.2	3.65	2.96	6.89	6.13	6.37	-0.34	5.01	-0.39	-0.23
12D→12C	-1.8	0.7	2.73	1.56	4.45	4.98	5.16	-1.13	4.02	-0.79	-0.60
13D→13C	0.1	0.1	4.31	3.20	7.81	6.59	5.16	0.02	4.06	0.81	1.09
14D→14C	0.3	0.3	1.64	0.84	2.32	2.84	2.31	-1.98	2.10	0.16	0.32
15A→15B	0.9	0.3	-0.62	2.65	2.23	1.02	4.63	6.55	5.22	0.02	0.01
15A→15C	-1.2	0.3	0.53	1.18	2.24	3.64	4.91	2.20	4.74	-1.87	-1.87
15B→15C	-2.2	0.3	1.15	-1.47	0.01	2.62	0.28	-4.34	-0.47	-1.88	-1.88
16A→16C	0.5	0.1	1.90	2.46	2.67	4.36	5.36	1.57	5.40	0.56	0.56

Questions also arise when having a look at the PCM/B3LYP calculations. Including electron correlation effects at a high-level should increase the predictive power of a model because the underlying physics are represented with high accuracy. This is the reason why a lot of researchers put work into the inclusion of electron correlation in DFT calculations.[79,80] However, the opposite is the case here, the results have worsened. Particular to this model is also the RMSE of the 5-membered rings, which is in the same range as the one of the 6-membered rings, and the MSE. The MSE here is the only one of all investigated models which is negative, the overall predicted values are smaller than for the other models, and there is a difference between 5- and 6-membered rings: 5-membered rings show a positive (0.55), 6-membered rings a negative MSE (-0.79). Besides, slopes are much higher and approach the critical value of one when the electron correlation is included. This leads to the assumption that by being the only model which yields the same accuracy for 5- and 6-membered rings and shows an overall good performance by only focusing at the errors, PCM/B3LYP without high-level electron correlation performs not well for physical reasons and thus is not recommended for tautomer predictions.

Comparison between the original computational approach during the SAMPL2 challenge from 2010 and the new one, developed during the SAMPL6 challenge, is worth discussing. This is also done in a recent publication by us.[10] The new approach includes a lot of physical improvements: Not only the inclusion of coupled cluster gas-phase energies and thermal corrections, but also the full QC electrostatic potential during 3D RISM calculations and the partial molar volume correction are applied. Assuming that a physically more accurate description of the molecule and the solvent improves the results, the new computational approach should lead to better results than the ones achieved in 2010. While being in the same range, the overall RMSE with the new workflow is worse with 2.2 compared to 2.0 kcal/mol, but more interesting is the look at the different ring sizes. In the original submission, there was an RMSE of only 0.57 kcal/mol for the 5-membered rings, which is more than 2 kcal/mol lower compared to the SAMPL6 approach (2.62 kcal/mol). The opposite is observed for the 6-membered rings: Here, the new approach is superior with an RMSE of 1.52 compared to 2.91 kcal/mol. Although the SAMPL2 approach results in a lower overall RMSE, the SAMPL6/CCSD(T) approach seems to be more reliable by having a fundamentally improved physical basis and a much smaller difference between 5- and 6-membered rings for which no obvious physical reason exists; it is the most consistent approach. Even the Lennard-Jones FF parameters used for 3D RISM calculations remained unchanged between both approaches, so this effect is not a FF artifact.

The inconsistency between the SAMPL2 and SAMPL6 setup regarding the predictive power for specific ring sizes gives rise to the question whether or not exists an experimental problem exists w.r.t the compounds 10-16. The partition function approach in SAMPL2 produced slightly worse results, quite in contrast to compounds 1-6 compared to the use of only the minimum energies, which is implausible and a hint for an experimental problem. The comparison with results from technically very different, though still QM-based models, here COSMO-RS data from Klamt and Diedenhofen[100] (they augmented hydration free energies with explicit gas phase calculations (at the MP2+vib-CT-BP-TZVP level of theory), similar to the SAMPL6/CCSD(T) approach), can help to get additional insight. They obtained an inverse trend compared to the original SAMPL2 submission, worse performance for the explanatory compared to the obscure set. In Figure 11, the juxtaposition between the two approaches is shown, together with the comparison between the original SAMPL2 and the SAMPL6 approach.

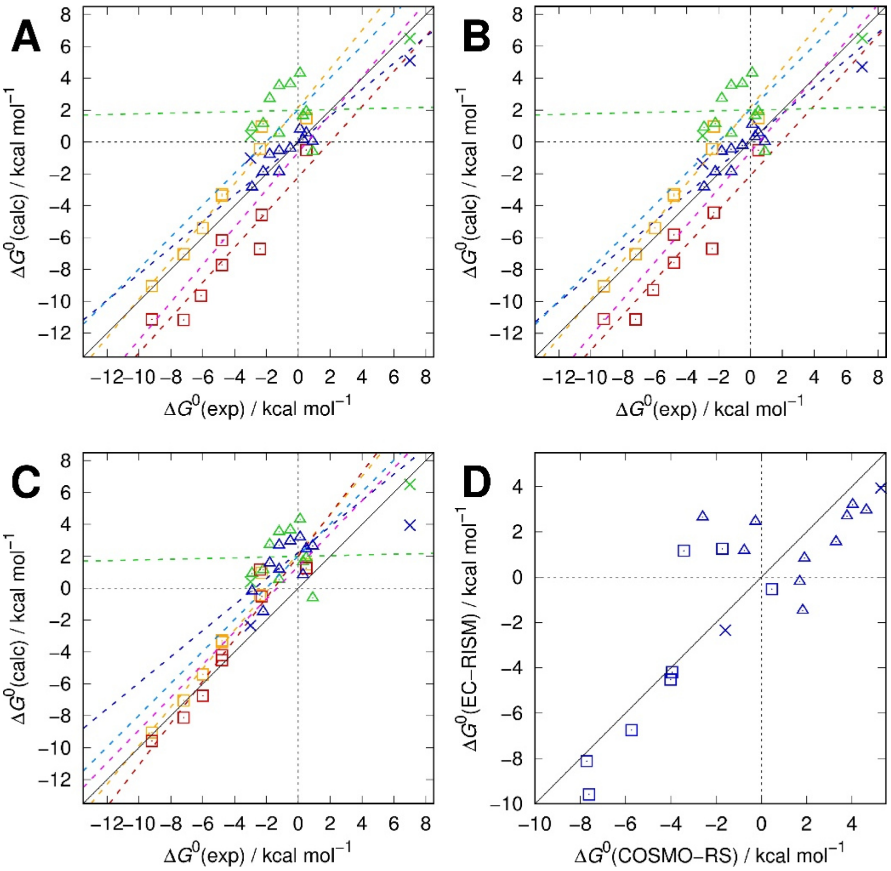


Figure 11: Calculated and experimental standard reaction Gibbs energies for the tautomer pairs of the SAMPL2 dataset (A–C)[90,91] and comparison of SAMPL6/CCSD(T) data with corresponding explicit COSMO-RS (MP2+vib-CT-BP-TZVP) results[100] (D). Data using the SAMPL6 workflow are shown as orange squares (obscure pairs 1-6), green triangles (explanatory pairs 10-16) and green crosses (explanatory pairs 7 and 8). Linear regressions are depicted as dashed lines in corresponding colors, with the total regression over all pairs in light blue (A-C). The data of the original SAMPL2 submission are shown by red squares (1-6), blue triangles (10-16) and blue crosses (7 and 8) with regression lines again in corresponding color, and total regression in magenta for the best performing SAMPL2 model (MP2/aug-cc-pVDZ/PSE-3) using only minimum conformations for SAMPL2 setup (A SAMPL2/ min and SAMPL6) or the Boltzmann weighted free energies of the conformational ensemble (B SAMPL2/Z and SAMPL6). Results from the explicit thermodynamic cycle combining SAMPL6-style Gibbs free energies of hydration and CCSD(T)/cc-pVTZ gas-phase free energies including B3LYP/6-311+G(d,p) thermal corrections are shown by analogously color-coded symbols in (C). Material from: Nicolas Tielker, Lukas Eberlein, Gerhard Hessler, K. Friedemann Schmidt, Stefan Güssregen, Stefan M. Kast, "Quantum–mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges?", Journal of Computer-Aided Molecular Design, published 2020, Springer.

The similarity, particularly for the strongly negative values, is obvious, while the data for the 5-membered rings is scattered more widely (RMSEs with respect to experiment of 2.62/3.82 for 10-16 and 1.52/1.50 kcal mol^{-1} for 1-6, when comparing SAMPL6/CCSD(T) and MP2+vib-CT-BP-TZVP, respectively).[10] This is a strong evidence that experimental reference data for the 6-membered rings are more reliable than for the 5-membered rings, even though the estimated experimental uncertainties are small. Averaging over both methods yields a hypothetical consensus prediction, the RMSEs of this consensus prediction relative to both individual predictions are smaller than the of each individual prediction to the experiment (1.07 (1-6), 1.25 (10-16), and 1.12 (1-16) kcal -mol^{-1}). This shows computational consistency, especially for the 5-membered rings which, together with the individual divergence from experiment, suggests that the experimental values for the 5-membered rings should be reconsidered. The pairs 7 and 8, which are neither 5- nor 6-membered rings are neglected in this discussion, since the experimental uncertainties are very high.

Focusing on the probably most well-known molecule pair within the dataset, the 2-pyridone/2-hydroxypyridine system (pair 1A/1B, see chapter 1.5), the computational predictions are in good agreement with the experimental value of -4.8 with -3.38 kcal/mol for the plain EC-RISM and -5.07 kcal/mol using the high-level gas-phase term. These results are more accurate than for all of the other methods benchmarked here (-2.83 (TI), -0.93 (PCM/MP2), -2.17 (PCM/MP2+CCSD(T)), -5.57 (PCM/B3LYP) -3.16 (PCM/B3LYP +CCSD(T)) and -7.73 (SAMPL2/min) kcal/mol) So, this system underlines the predictive quality of the newly developed EC-RISM approach for tautomeric phenomena.

Further, the improvements made for the SAMPL6 computational approach should allow the calculation of energy independent observables, like NMR chemical shifts, with a higher accuracy by using the MP2 level of theory and full QC electrostatic potential.[89] This is investigated in the following chapters.

4.3 Probing the sensitivity of chemical shifts for tautomer predictions - histamine

In the previous chapters, the EC-RISM solvation model was extended for the calculation of various temperature conditions, and the predictivity of EC-RISM for tautomer ratios is benchmarked. Together with the PMV correction for high-pressures, it is possible to calculate the tautomer ratios of nucleobases pressure- and temperature-dependent. But larger systems than the rigid nucleobases with little conformational freedom, only a few rotable bonds on an aromatic scaffold, like nucleotides, are far more complicated. For these systems, an exhaustive conformer generation followed by EC-RISM calculations is not possible, it would need a too large computational effort, especially when NMR calculations have to be performed. Therefore, benchmarking on a well-known test-system with more conformational and tautomeric freedom in multiple ionization states is needed. This is done in this chapter on the test system histamine. Additionally, experimental NMR chemical shifts are known for the histamine (measured by Wolf Hiller), which allows for developing and benchmarking of a combined computational and NMR spectroscopic approach for tautomer predictions. The benchmark system histamine is well known and was also investigated in the authors master's thesis,[277] but is presented here in more detail and on a new computational basis; some of the workflows used there are similar to those used here, but no results have been transferred.

Histamine is a small neurotransmitter involved in the regulation of multiple physiological functions like sleep-wake regulation, gastric acid release and schizophrenia, but most familiar in immune responses. Histamine has three protonation states: the free base and the monocation, the dominant species at physiological pH, at which several tautomers are present, and the fully protonated dication. The protonation and tautomerism take place at the nitrogen atoms of histamine, especially the π- and τ-nitrogen (Figure 12). In the following, the different tautomers will be named according to their protonated nitrogen atoms. The experimental data is already used in the authors master's thesis.[277] The resulting chemical shifts and coupling constants (the latter not used here, since no coupling constants can be calculated at the MP2 level of theory with the Gaussian program package), assigned to the respective atoms, are shown in the appendix in Table 50 with the nomenclature depicted in Figure 12.

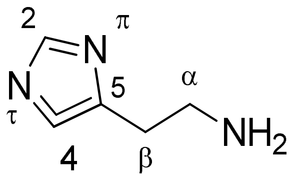


Figure 12: Nomenclature of histamine; the different tautomers are named by their protonated nitrogen atoms, and the nitrogen atom next to the α carbon is called α-nitrogen. If multiple nitrogen atoms are protonated, they are separated by a "+" symbol (for example α+π-tautomer). The rotable bonds which are rotated for the exhaustive conformational sampling are the three of the side chain between carbon 5 and the α nitrogen.

Previous studies that focused on the analysis of this tautomerism exhibit diverse results (Table 11). In this work, six histamine tautomers are examined; the π- and τ-tautomer of the free base (a zwitterionic α-tautomer is chemically implausible), the α+π-, α+τ- and π+τ-tautomers of the monocation, and the dicationic α+π+τ-tautomer. These tautomers and the number of conformations for each tautomer, from an exhaustive search, are shown in Figure 13.

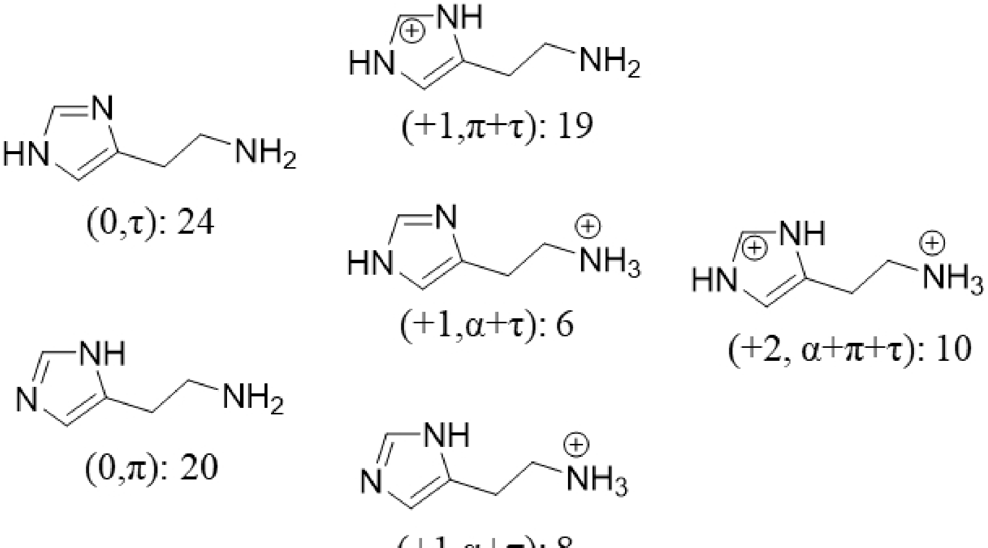


Figure 13: Investigated tautomers of histamine. The tautomers of the free base are shown on the left, the ones of the monocation in the middle, and the one of dication on the right side. Additionally, the number of conformations of each tautomer is given.

The experimental pK_a values are 9.75 for the transition from the free base to the monocation and 6.04 for the shift to the fully protonated dication. Using EC-RISM, the p*K*a value can be calculated with Eqn. 186, which was successfully parametrized and used in the SAMPL6 blind prediction challenge. This way, the resulting pK_a values are 7.53 and 2.71 (Table 10); they show a low correlation to the experiment and the predictive quality of the model is worse than suggested within the SAMPL6 challenge.[4] This is an important development since the master's thesis in which no pK_a-model was applied for the determination of pK_a-values; instead, in this work, the solvation free energy of the proton was fitted using the experimental histamine pK_a-values. The details of this approach can be reviewed in Ref. 18 and 277, but an important problem is that a single parameter, the solvation free energy of the proton, is fitted using two points (both experimental pK_a-values).

The protonation of histamine takes place at the nitrogen atoms, which leads to the reasonable assumption that the nitrogen Lennard-Jones parameters contribute to the performance issues with the SAMPL6 approach, not only for the pK_a-values here, it may also play an important role for the SAMPL2 tautomer pairs. Since these nuclei are also the most important ones for the NMR calculations, it is important that the nitrogen atoms are described with high accuracy. In the GAFF force field, which is used for EC-RISM, there are multiple nitrogen atom types, for example for aromatic nitrogen atoms or nitrogen atoms with a specific number of binding partners. A further investigation showed that the calculated Gibbs free energy of hydration for the imidazole, a histamine without side chain and part the MNSOL database, is in good agreement with the experimental value of 2.12 kcal/mol. In contrast, the neutral primary amines, analogues to the histamine side chain and, like the side chain all having the n3 (sp^3 nitrogen atom with three binding partners) Lennard-Jones parameter, in the MNSOL show a large deviation from the experiment with an RMSE of 4.25 kcal/mol.[4,247] The n4 parameter for sp^3 nitrogen atoms with four binding partners, as the protonated histamine side chain, has the same LJ-parameters in the original GAFF even though the n3 is expected to need a larger σ, since it should be a larger atom due to the free electron pair. Therefore, a reparametrization of this n3 atomtype was done, using the primary amines from the MNSOL; the molecules used and the results of the reparametrization are given in *4.3.1*, the workflow is described in 3.2.3.1. The structures of histamine and the dataset used for the n3 atomtype reparametrization as well as all FF parameters and NMR chemical shifts are given in the SI part 03.

4.3.1 Reparametrization of GAFF parameters

The MNSOL was manually searched for primary amines; the molecules together with the codes used in the MNSOL, the experimental and calculated results, the latter using the original GAFF and with the reoptimized n3 atomtype (called GAFF(n3)), are given in Table 9. The results of the reparametrization and trends during the fitting process are shown in Figure 14.

Table 9: MNSOL codes and molecule names of the reparametrization dataset. The dataset contains the neutral primary amines from the MNSOL for which experimental solvation free energies in water are available. The experimental and calculated values are given, the latter using the original GAFF and the reparametrized n3 atomtype (GAFF/n3).

Code	Molecule	Exp.	Calc. (GAFF)	Calc. (GAFF/n3))
0103eth	Ethylamine	-4.50	-7.09	-4.24
0106pro	Propylamine	-4.39	-7.01	-4.22
0110but	Butylamine	-4.29	-6.75	-3.97
0113pen	Pentylamine	-4.10	-6.53	-3.73
0147met	2-methoxy-ethanamine	-6.55	-9.02	-7.34
0228met	Methylamine	-4.56	-7.45	-3.98
0229hyd	Hydrazine	-6.26	-13.15	-7.94
n005	Methyl-hydrazine	-5.31	-11.80	-7.38
n006	1,1-dimethyl-hydrazine	-4.48	-10.05	-6.21
n016	1,2-ethane-diamine	-9.72	-13.65	-7.94

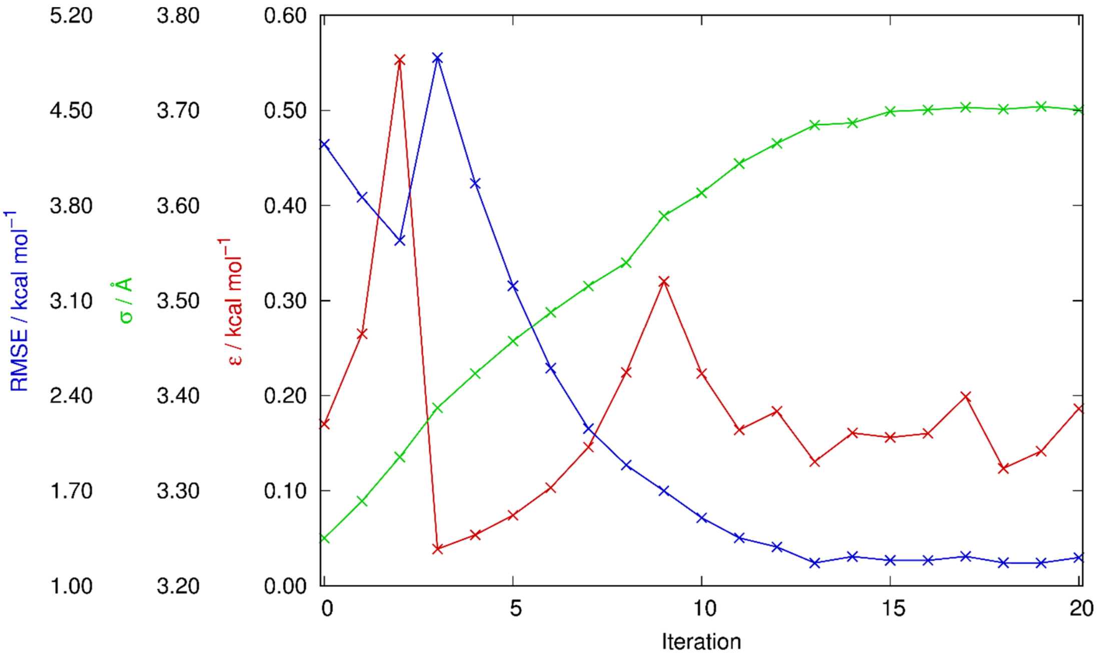


Figure 14: Results of the reparametrization process of the n3 parameter of the GAFF force field. Starting with the original parameters and RMSE, the trends of σ (green, in Å), ε (red) and RMSE (blue, both in kcal/mol) are shown until convergence.

The original GAFF values are 3.25 Å for the σ and 0.17 kcal/mol for ε, yielding an RMSE of 4.25 kcal/mol for the MNSOL primary amines. During the fitting process, the overall trend is an increasing σ, coherent with a declining RMSE. In the first steps, ε is increasing with a large step from iteration 2 to 3 (approximately 0.28 to 0.55 kcal/mol), while the RMSE is declining. In the next step, iteration 4, it rapidly declines to less than 0.05 kcal/mol, pushing the RMSE to a level higher than for the original GAFF parameters. This is followed by an increase of ε to values again larger than the original parameter, and leads to an oscillation around this parameter. After the size of the ε exceeds a specific level, and during the oscillation around the original parameter, it seems to have little impact on the RMSE, which shows an inverse correlation to the size of σ and therefore the n3 nitrogen atoms. As long as no extreme values of ε occur, the system seems to be dominated by the influence of the σ parameter, which is too small for a sufficient representation of the n3 nitrogen atoms in the original force field. This size dependence is plausible, considering that the same set of LJ-parameters is used in the GAFF for the n3 and n4 atomtypes and that the n3 has a free electron pair, and, thus, should be larger than the n4. In the GAFF, all nitrogen atomtypes have the same parameter set, which is a rough approximation since the size of the nitrogen strongly depends on the existence of a free electron pair and the existence of strong π-interactions in an aromatic system. Therefore, it is possible that a reparametrization of more nitrogen atomtypes is needed to account for the fine differences in chemical properties of the nitrogen atoms in their specific environment.

Repeating the histamine calculations for the tautomers with an n3 atomtype, which is generally called n3 or EC-RISM/n3 in this chapter, yields a new set of computational pK_a values of 9.29 and 3.28. While the transition of the free base to the monocation is in good agreement with the experiment with an error of only 0.46 pK_a-units, open questions remain regarding the transition to the dication. The deviation to the experiment is reduced by 0.57 pK_a-units but still large with 2.76 pK_a-units. A reparametrization of the n4 parameter, which is part of the histamine tautomers with a protonated side-chain, could help to overcome this large difference, but is not reasonable since the error in the computed Gibbs free energies of hydration for the MNSOL compounds with this parameter is much smaller than the one of the n3 compounds with an RMSE of 2.22 kcal/mol. A possible explanation is the underlying pK_a model, which was parameterized with respect to transitions between neutrals, monocations and monoanions, while the histamine transition is one to a dication. Dications were also part of the SAMPL6 and SAMPL7 datasets and did not cause any issues there, but the histamine is a much smaller molecule than the SAMPL6 and SAMPL7 compounds, which may be problematic. The pH-dependent ionization state fractions are shown in Figure 15, the curves of the free base fraction are shifted from the experimental ones even with the reparametrization, but the fractions of the mono- and dication fit the experimental data quite well.

Due to the time sequence in which the calculations were carried out, it was unfortunately not possible to use the new n3 atomtype for the calculations on the nucleobases and nucleotides in chapter 4.5 to 4.7. The n3 atomtype is not present in the aromatic backbone of the nucleobases and plays only a role in some tautomers with an amino-group. It is therefore expected to have a small influence on the calculations of the nucleobases and nucleotides. Test calculations on the adenine prove that; the new n3 atomtype changes the reaction free energies from the Watson-Crick tautomer to the minor N1- and N3-tautomers (in detail explained in the respective chapters) only by 0.00003 kcal/mol (N1-tautomer), respectively 0.00011 kcal/mol (N3-tautomer).

Table 10: pK_a values of histamine calculated via Eqn. 186 using EC-RISM and EC-RISM with the reparametrized GAFF atomtype n3 (Lennard-Jones parameters σ=3.70055 Å and ε=0.18627 kcal/mol (EC-RISM(n3))). (Original parameters: 3.25 Å and 0.17 kcal/mol)

	p$K_{a,1}$	p$K_{a,2}$
Experiment[273]	6.04	9.75
EC-RISM	2.71	7.53
EC-RISM(n3)	3.28	9.29

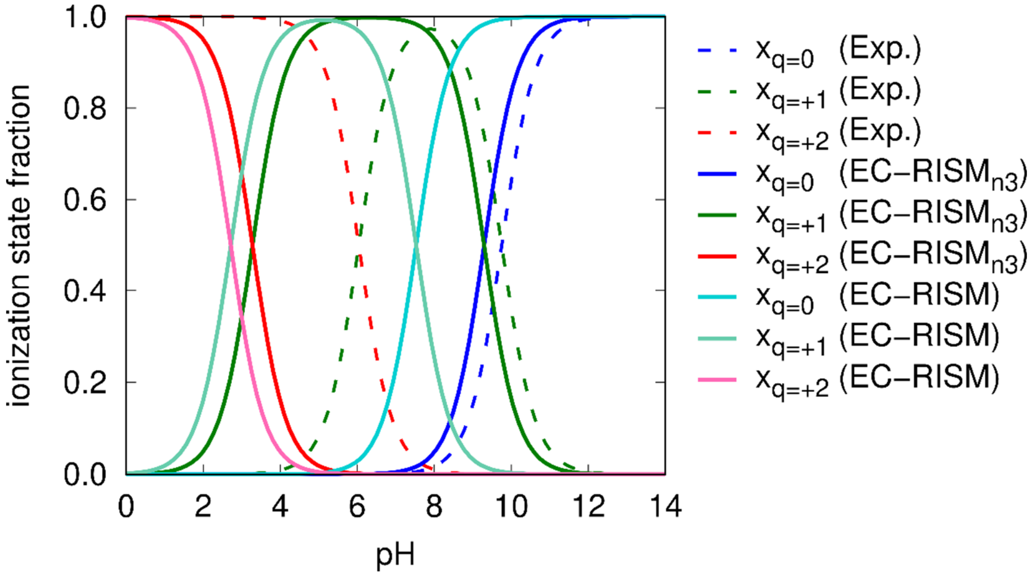


Figure 15: pH-dependent ionization state fractions of histamine based on experimental[273] and calculated pK_a values. The calculations are done with EC-RISM without and with (subscript n3) reparametrization of the n3 GAFF atomtype.

4.3.2 Calculation of histamine tautomerism

With the reparametrized n3 atomtype, the pK_a-values of histamine, especially for the transition from the free base to the monocation, could be improved. With the reparametrized GAFF/n3 and the original GAFF, the tautomerism of histamine is investigated computationally. Previous studies did not take the π+τ-tautomer of the monocation into account, which was correct as the energy calculations indicate. The monocation is dominated by the α+τ-tautomer with fractions of over 90% (nearly 100% by including electron correlation effects at a high level using CCSD(T) calculations), the α+π-tautomer is the minor tautomer with a low fraction, and the π+τ-tautomer is not populated; they are penalized by 1.74 and 5.92 kcal/mol, respectively. The differences between the use of original GAFF or the reparametrized n3 parameter is negligible, since it changes only the energy of the not significantly populated π+τ-tautomer. The free base main tautomer is the τ-tautomer with a fraction of 55-63%, but the π-tautomer has a strong contribution, and the energetic differences between both tautomers are small. The reparametrization of the n3 GAFF parameters even lowers the energetic difference. Including CCSD(T) data, the π-fraction is lower than without while the original GAFF is used; it does not change the tautomer fractions by using the reparametrized GAFF, this underlines the robustness of the method. The resulting pH-dependent tautomer and ionization state fractions of the EC-RISM/n3 calculations are shown in Figure 16 together with the conformer distribution.

Table 11: Tautomer fractions of histamine. Calculated via EC-RISM with (marked by the subscript n3) and without the reparametrized n3 GAFF atomtype, in combination with (subscript CCSD(T)) and without the explicit high-level gas-phase energies, as well as PCM results and data from the literature. The tautomer fractions within each ionization state add to 1. In addition, the reaction free energies (in kcal mol^{-1}) to the respective tautomers (from the τ- (free base), respectively α+τ-tautomer (monocation)) are given.

	x_π	x_τ	$x_{\alpha+\pi}$	$x_{\pi+\tau}$	$x_{\alpha+\tau}$	$\Delta_r G(\pi)$	$\Delta_r G(\alpha+\pi)$	$\Delta_r G(\pi+\tau)$
EC-RISM	0.37	0.63	0.08	0.00	0.92	0.32	1.49	5.55
EC-RISM$_{\text{n3}}$	0.45	0.55	0.05	0.00	0.95	0.12	1.75	5.92
EC-RISM$_{\text{CCSD(T)}}$	0.30	0.70	0.01	0.00	0.99	0.49	2.69	6.40
EC-RISM$_{\text{n3,CCSD(T)}}$	0.45	0.55	0.00	0.00	1.00	0.12	3.25	6.04
PCM	0.54	0.46	0.01	0.09	0.90	-0.09	3.01	1.36
PCM$_{\text{CCSD(T)}}$	0.56	0.44	0.00	0.03	0.97	-0.13	4.98	1.99
Reynolds *et al.*[291]	0.20	0.80	0.20	-	0.80	0.82	0.82	-
Wasylishen *et al.*[292]	0.20	0.80	0.20	-	0.80	0.82	0.82	-
Nagy et al.[293]	0.17	0.83	0.34	-	0.66	0.94	0.39	-
Forti et al.[294]	0.48	0.52	-	-	-	0.05	-	-

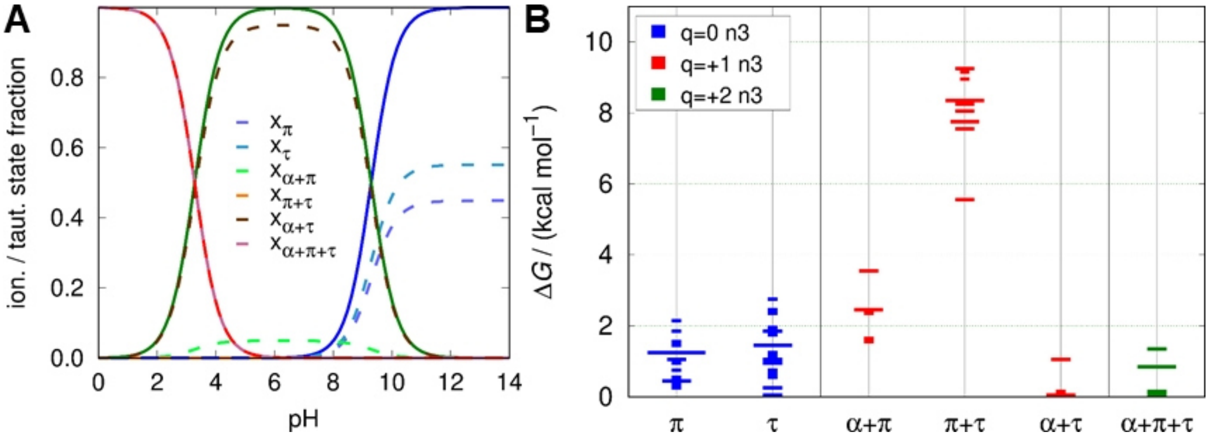


Figure 16: Computed pH-dependent ionization state (solid) and tautomer (dashed) fractions calculated using EC-RISM$_{n3}$ (A) and the corresponding energetic distribution of conformers for all tautomers and ionization states (B). The baseline is obtained by taking the minimum energy of the respective ionization state; the width of the bars indicates the number of conformations in this energetic region.

The small energetic difference between π- and τ-tautomers can be explained by the broad distribution of many conformations of both tautomers in a small energy range, while the α+τ-conformations are all energetically lower than the α+π-conformations, which are again lower than the π+τ-conformations. Using PCM, the main tautomer of the free base and the minor tautomer of the monocation switch compared to EC-RISM; the π-tautomer is the free base main tautomer with fractions of 0.54 (PCM) and 0.56 (PCM/CCSD(T)), and the π+τ-tautomer is the minor monocation tautomer (fractions of 0.09 with PCM and 0.03 with PCM/CCSD(T)), which is not in agreement with EC-RISM and former experimental and computational results. This can clearly be seen by the distribution of conformations, shown in Figure 17 for the remaining levels of theory: the minimum conformation of the free base is a π-conformation, but still the distribution is in a small energy range while the π+τ-conformations are much closer and partially overlapping with the energetic range of the α+τ-conformations; the α+π-conformations are penalized by approximately 2 kcal/mol stronger compared to EC-RISM/n3. The conformer distributions are more spread using CCSD(T) gas-phase energies, there are fewer conformations within the same energetic range, so it seems necessary to include this data to distinguish between the conformations correctly. However, the overall trends in tautomer populations are not affected by this since the inclusion does not change the energetic order of the minimum conformations. The EC-RISM results, especially with the reparametrized n3 atomtype, are in close agreement with the results from Forti *et al.*[294] but deviate stronger from the other literature data than using the original GAFF. The experimental data from Reynolds *et al.*[291] and Wasylishen *et al.*[292] are based on NMR spectroscopy, chemical shifts and coupling constants, respectively, on methylated histamine derivatives, while Nagy *et al.*[293] and Forti *et al.* present computational results. Since the approach of Forti *et al.* is done using conformational sampling and state of the art QC methods, it seems, together with the EC-RISM results, to give the most reliable results. The via the n3 atomtype reparametrization increased agreement with the literature (Table 11) indicated, that the underlying conformational fraction with this method are reliable. Nevertheless, in the next chapter an additional consistency check is presented in the form of NMR calculations.

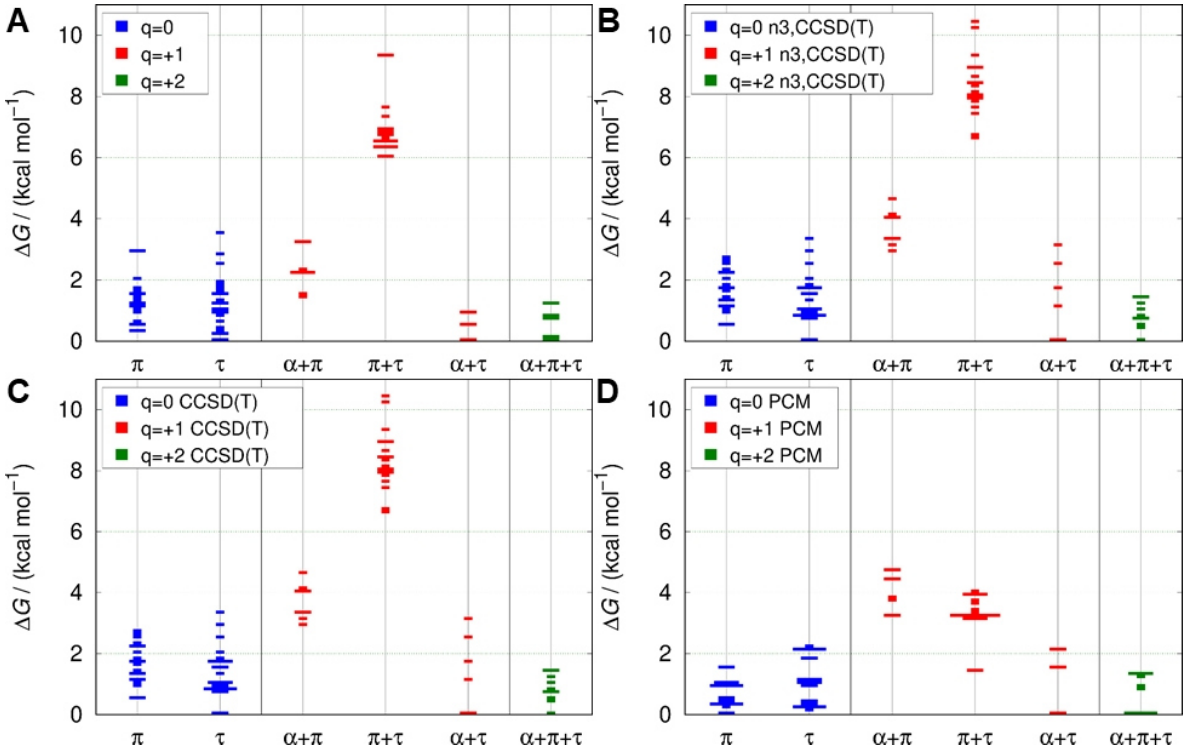


Figure 17: Energetic distribution of conformers for all tautomers and ionization states. The baseline is obtained by taking the minimum energy of the respective ionization state; the width of the bars indicates the number of conformations in this energetic region. The results are obtained at the EC-RISM (A), EC-RISM/CCSD(T) (C) EC-RISM$_{n3}$/CCSD(T) (B) and PCM (D) level of theory.

4.3.3 Calculation of histamine NMR chemical shifts

Since the energetics of histamine are clarified, another observable, NMR chemical shifts, is investigated. The NMR chemical shifts are an QC observable which is only indirect related to the energetics of a molecule and therefore allow an additional consistency check. Furthermore, the chemical shifts can be calculated for each conformation, in contrast to the experiment, in which only an ensemble average is measured for systems with fast transitions like the histamine, this allows for the fitting of tautomer populations as an additional way to determine tautomer fractions. This also helps to verify the validity of the reparametrization approach described above. The experimental NMR chemical shifts were measured by Wolf Hiller and given in the appendix in Table 50. Experimental data were recorded for all nuclei in histamine, ^{1}H, ^{13}C and ^{15}N (since the nitrogen chemical shifts are referenced to nitromethane in D_2O, a value of 383.87 ppm has been added as secondary reference to correct to liquid ammonia as reference[295]), so calculations were also done for all of these kinds of nuclei. To calculate chemical shifts, the shielding constants of the reference nuclei are needed. Here, reference shielding constants of 278.087 (^{15}N), 200.934 (^{13}C), and 31.8997 ppm (^{1}H) are used;[7] since the nitrogen reference shielding is calculated for nitrogen in water instead (according to the IUPAC) liquid ammonia, a secondary reference of -19.4 ppm[296] has to be added. A detailed discussion about computational referencing methods and the development of ambient condition as well as temperature and pressure dependent reference shielding constants is given in chapter 4.4.

In this chapter, the correlation between experimental results for the respective ionization states and the computed chemical shifts for the ionization states, tautomers and main conformation of each ionization state is investigated. Since the calculated conformational and tautomeric fractions are needed to calculate the chemical shifts of the tautomers and ionization states respectively, this allows not only to judge the accuracy of the chemical shift calculations; also the quality of these underlying fractions can be investigated. The results of this comparison are summarized in Table 12 and depicted in Figure 18 (ionization states), Figure 19 (tautomers), and Figure 20 (main conformations).

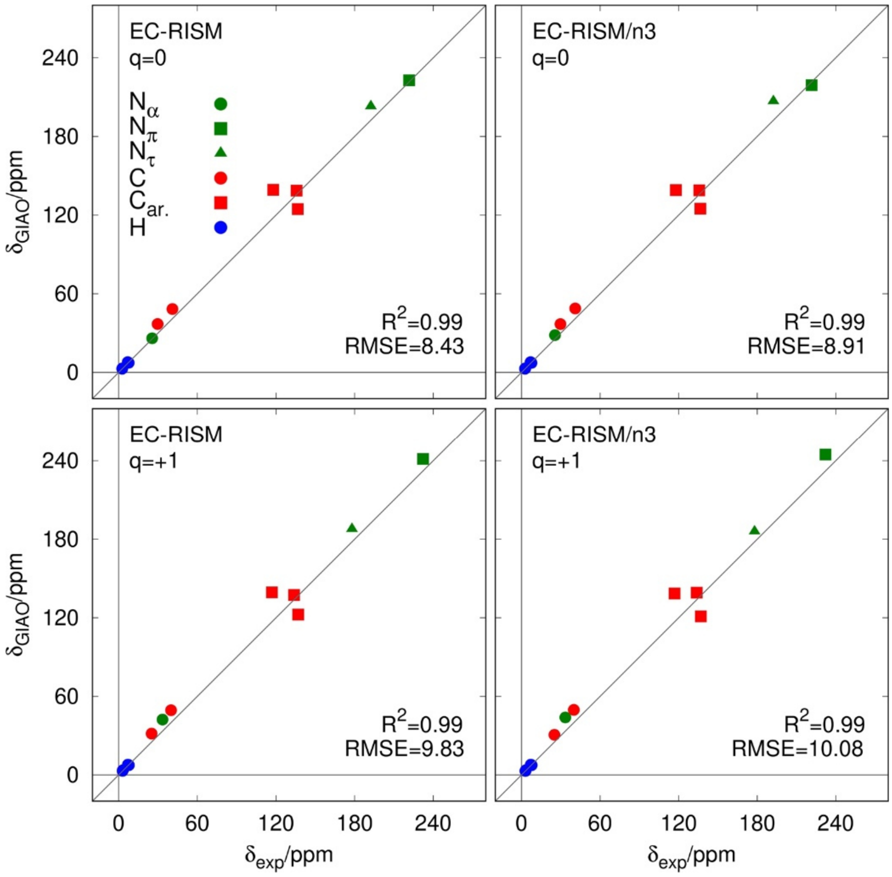


Figure 18: Calculated vs. experimental chemical shifts of the histamine ionization states. Calculations are done with (EC-RISM/n3) and without (EC-RISM) the reparametrized n3 atomtype. In addition, the RMSE between experiment and calculation, and the coefficient of determination are given.

The comparison of computed and experimental NMR chemical shifts of the respective ionization states results in similar RMSEs for EC-RISM and EC-RISM/n3 for ionization states, with a slightly lower one using the original GAFF force field. This shows that, even though the reparametrization of the n3 atomtype influenced the quality of pK_a-predictions drastically, it does not significantly influence the quality of NMR chemical shift predictions. This shows that the FF parameters are not influencing the calculation of NMR chemical shifts in a way they influence the energy calculations; the error of the NMR calculations seems to be dominated by the QC calculation. The small differences can be caused either by the small changes in the conformer distribution (Figure 16 and Figure 17) or by the influence of the different solvent distribution on the wave function.

The RMSE of the τ-tautomer is much smaller than of the π-tautomer (approximately 11 to 17 ppm). According to the calculated tautomer fractions, the RMSE of the α+τ-tautomer is also lower than the one of the α+π-tautomer with an even larger difference (approximately 10 to 23.5 ppm), this is consistent to the EC-RISM calculations and the literature. Surprisingly, the RMSE of the π+τ-tautomer is in the same range as the α+τ-RMSE with 14.47 (GAFF) and 12.33 ppm (n3). This may be a coincidence or be caused by the intrinsic error of the QC calculations. It is noteworthy that the RMSE of the π+τ-tautomer is smaller at the level of theory for which the calculated fraction is also smaller (0.000078 with GAFF and 0.000043 with n3). This is an inconsistency since a disfavorisation of this tautomer in both observables could be expected from the reparametrization of the n3 atomtype Lennard-Jones parameters, since the reparametrization leads to an increased accuracy for the pK_a calculations. This hints that the RMSE of the NMR chemical shifts is a coincidence and shows how important the calculation of two independent observables is to verify results.

Table 12: Error metrics he RMSE, MSE and MAE (in ppm) of the NMR parameters calculated for the tautomers and ionization states to the respective corresponding experimental data and the linear regression data slope m, intercept b in ppm and coefficient of determination (R^2) are shown. It is important to notice, that the calculation of the tautomer and ionization state NMR chemical shifts is done using the calculated fractions of the underlying species.

Tautomer	π	τ	α+π	π+τ	α+τ
RMSE/MAE/MSE	17.73/10.37/3.35	11.08/7.83/3.52	23.52/13.58/4.50	14.47/9.09/-1.50	10.14/7.51/4.80
$m/b/R^2$	0.95/0.78/0.95	0.96/-0.31/0.98	0.93/1.32/0.91	1.07/-3.51/0.97	0.96/-1.75/0.99
RMSE/MAE/MSE$_{n3}$	17.36/10.34/3.63	11.17/8.10/3.74	23.53/13.58/4.50	12.33/7.67/-1.00	10.36/7.71/4.89
$m/b/R^2{}_{n3}$	0.95/0.33/0.95	0.96/-0.68/0.98	0.93/1.34/0.91	1.06/-3.56/0.98	0.96/-1.64/0.99
Main conformation	π	τ	α+π	π+τ	α+τ
RMSE/MAE/MSE	17.79/10.35/3.25	10.97/7.73/3.46	23.64/13.56/4.48	14.96/9.41/-1.74	9.75/7.11/4.63
$m/b/R^2$	0.95/0.86/0.95	0.96/-0.33/0.98	0.93/1.41/0.91	1.07/-3.21/0.97	0.97/-1.83/0.99
RMSE/MAE/MSE$_{n3}$	17.78/10.51/3.50	11.04/7.92/3.70	23.64/13.56/4.48	12.07/7.54/-0.94	9.75/7.11/4.63
$m/b/R^2{}_{n3}$	0.95/0.44/0.95	0.96/-0.73/0.98	0.93/1.41/0.91	1.06/-3.50/0.98	0.97/-1.83/0.99

Ionization state	Free base	Monocation
RMSE/MAE/MSE	8.43/5.62/3.46	9.83/7.36/4.78
$m/b/R^2$	0.97/-1.46/0.99	0.97/-2.03/0.99
RMSE/MAE/MSE$_{n3}$	8.91/6.12/3.69	10.08/7.60/4.87
$m/b/R^2{}_{n3}$	0.98/-1.81/0.99	0.96/-1.84/0.99

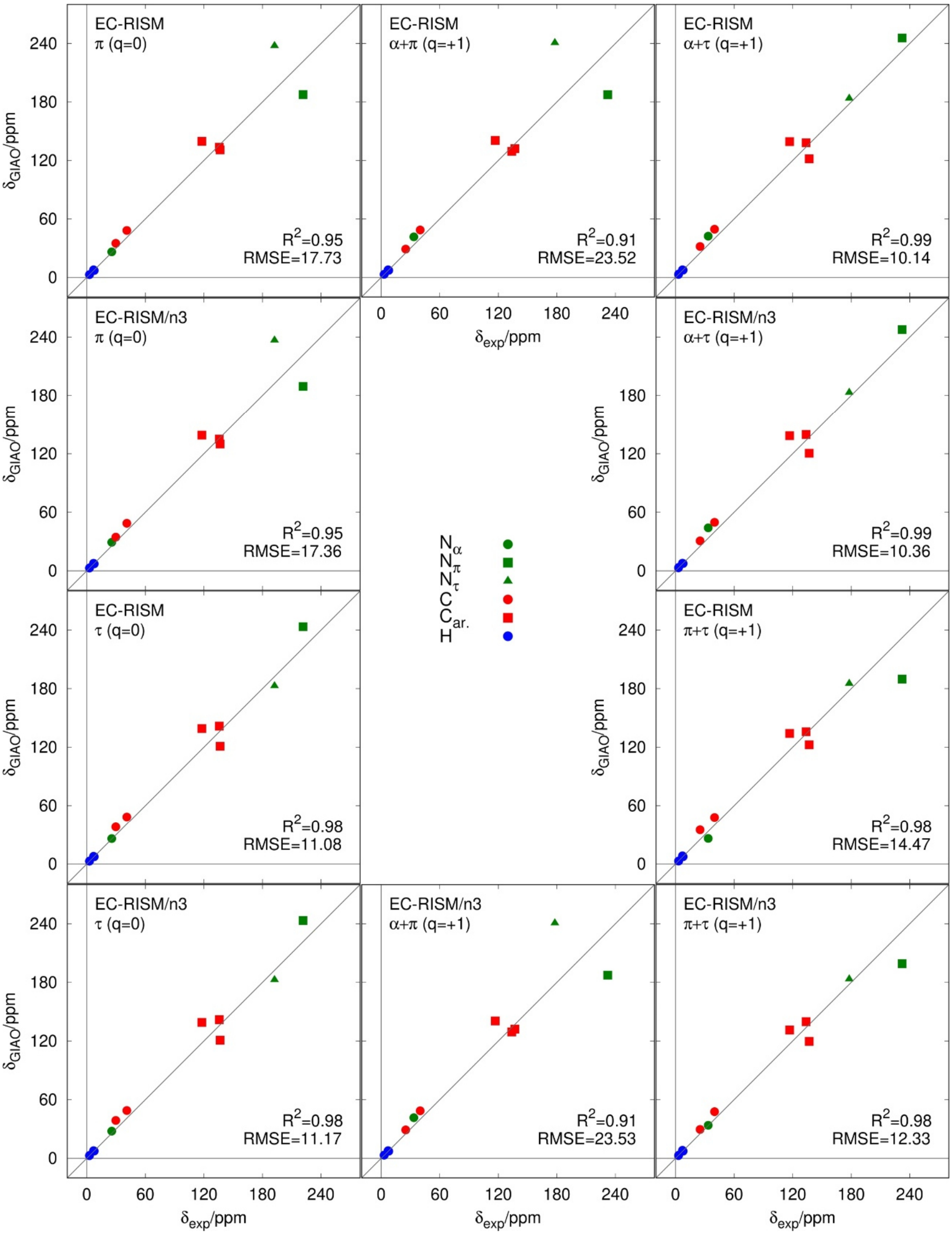


Figure 19: Calculated vs. experimental chemical shifts of the histamine tautomers. Calculations are done with (EC-RISM/n3) and without (EC-RISM) the reparametrized n3 atomtype. In addition, the RMSE between experiment and calculation, and the coefficient of determination are given.

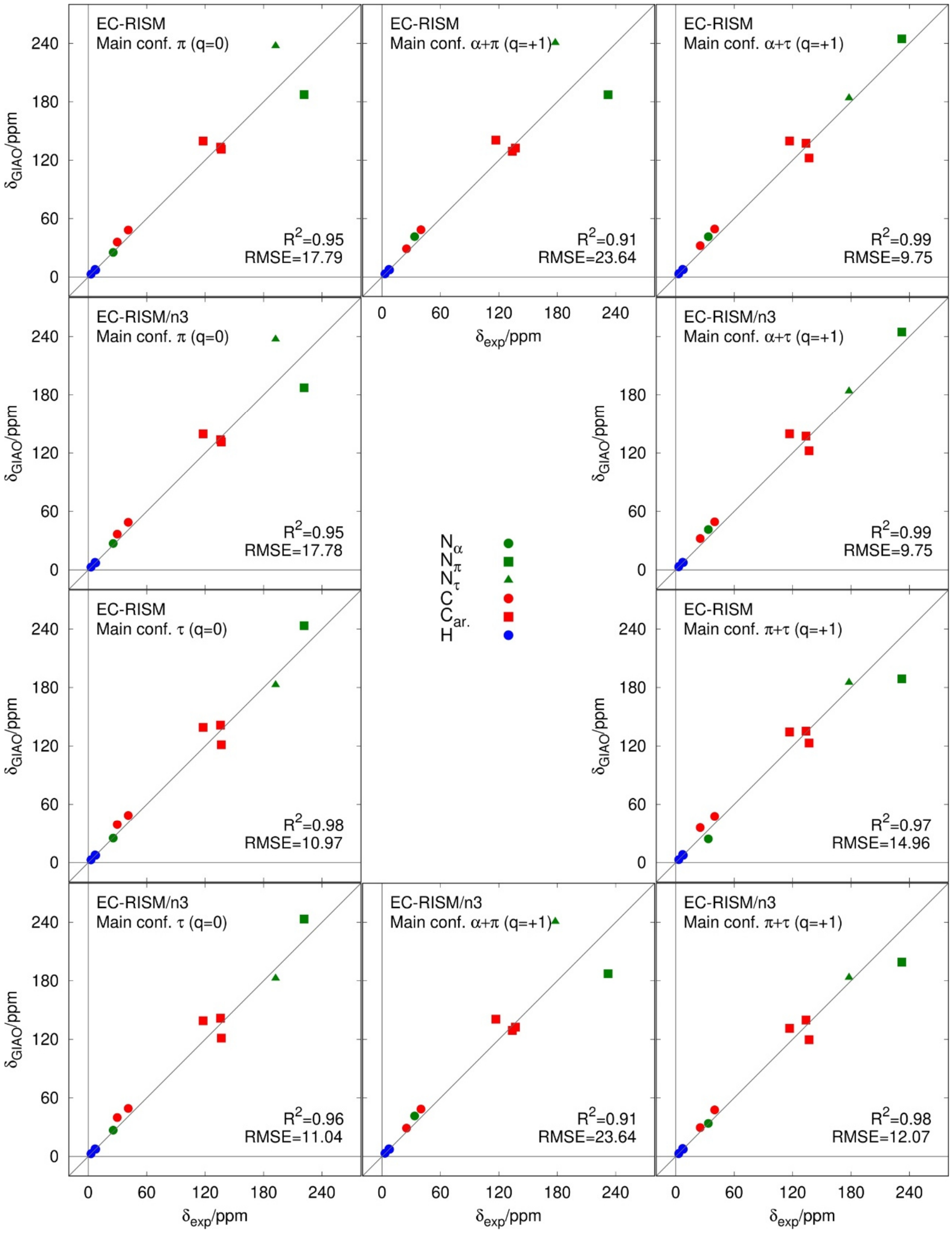


Figure 20: Calculated vs. experimental chemical shifts of the histamine main conformations. Calculations are done with (EC-RISM/n3) and without (EC-RISM) the reparametrized n3 atomtype. In addition, the RMSE between experiment and calculation and the coefficient of determination are given.

The differences between the deviation of the calculated chemical shifts of the tautomers and the main conformations from the experiment are small. There is a consistency in the correlation between the main conformations, tautomers and ionization states, the errors for the ionization states are smaller (free base) or in the same range (monocation) than for the main tautomers, and the errors of the tautomers are in the same range as those of the main conformations. This hints, together with the energy calculations, which are consistent to newer literature data and the pK_a-values, which are in good agreement with the experimental values (neglecting the influence of the double charged dication with only a single tautomer), that the underlying fractions are valid.

4.3.4 Extraction of histamine tautomer fractions from NMR chemical shifts

In the past chapters, the investigation of the quality of histamine energetics calculations and the agreement between calculated and experimental chemical shifts was done. A further possibility to investigate the histamine tautomerism is the use of the calculated NMR chemical shifts by relying as less as possible on the energy calculations. The chosen way is to calculate the chemical shifts of the tautomers and fit them to the experimental shifts of the respective ionization states using Eqn. 196. This way, the calculated energetics of the conformations are still included in the fitting process since they are needed to calculate the chemical shifts of the tautomers, but the energetic differences between the tautomers are neglected. The calculated chemical shifts of the tautomers and the experimental shifts of the ionization state are used directly in the fitting process and normalized using Eqn. 197, a normalization using the mean value of all chemical shifts of the respective nuclei, hereafter called norm$_N$, and Eqn. 198 using an additional normalization to the number of the respective nuclei, called norm$_a$. The fitting is done using all three kinds of chemical shifts, with and without normalization, and all possible combinations of nuclei (all nuclei, only ^{1}H, ^{13}C and ^{15}N respectively) and combinations of ^{1}H/^{13}C, ^{1}H/^{15}N and ^{13}C/^{15}N chemical shifts. This way, not only the tautomer ratios can be calculated, but also the sensitivity of the respective nuclei for changes in the tautomeric state of a molecule can be investigated. This is done with NMR chemical shifts calculated via EC-RISM and EC-RISM/n3. Additionally, the fitting for the monocation is done with and without inclusion of data of the π+τ-tautomer since the energetics calculations suggest it is non-existing in solution. The results of the fitting process are given in Table 13. The raw data, including the chemical shifts used for the fitting process, are given in the SI part 03.

Table 13: Results of the calculation of tautomer fractions from experimental NMR chemical shifts. The results are obtained by using all chemical shifts or only the ones of specific types of nuclei which are given as subscripts. The computational chemical shifts of the tautomers used in the fitting process are calculated via EC-RISM with (marked by the subscript n3) and without the reparametrized n3 GAFF atomtype. The first value in each column is from calculations using the chemical shifts directly, the second by using the normalization given in Eqn. 197 for calculated and experimental chemical shifts, and the third using the Eqn. 198 for the normalization. For the monocation tautomers, there are two rows of numbers, the top is calculated considering each monocation tautomer as possible, the bottom by neglecting the $\pi+\tau$-tautomer. The fractions of each tautomer add up to 1 within each ionization state.

	x_{π}	x_{τ}	$x_{\alpha+\pi}$	$x_{\pi+\tau}$	$x_{\alpha+\tau}$
EC-RISM	0.37	0.63	0.08	0.00	0.92
EC-RISM$_{n3}$	0.45	0.55	0.05	0.00	0.95
EC-RISM$_{CCSD(T)}$	0.30	0.70	0.01	0.00	0.99
EC-RISM$_{n3,CCSD(T)}$	0.45	0.55	0.00	0.00	1.00
Forti *et al*.[294]	0.48	0.52	-	-	-
Method	x_{π}	x_{τ}	$x_{\alpha+\pi}$	$x_{\pi+\tau}$	$x_{\alpha+\tau}$
NMR(δ)	0.32/0.33/0.29	0.68/0.67/0.71	0.00/0.09/0.00	0.31/0.05/0.14	0.69/0.86/0.86
			0.11/0.11/0.07	-	0.89/0.89/0.93
NMR$_{n3}$(δ)	0.32/0.34/0.30	0.68/0.66/0.70	0.00/0.09/0.03	0.40/0.11/0.13	0.60/0.80/0.84
			0.12/0.13/0.08	-	0.88/0.87/0.92
NMR(δ_{H})	0.00/0.00/0.00	1.00/1.00/1.00	0.00/0.00/0.00	0.00/0.00/0.00	1.00/1.00/1.00
			0.00/0.00/0.00	-	1.00/1.00/1.00
NMR$_{n3}$(δ_{H})	0.00/0.00/0.00	1.00/1.00/1.00	0.00/0.00/0.00	0.00/0.00/0.00	1.00/1.00/1.00
			0.00/0.00/0.00	-	1.00/1.00/1.00
NMR(δ_{C})	1.00/1.00/1.00	0.00/0.00/0.00	0.47/0.72/0.72	0.53/0.28/0.27	0.00/0.00/0.01
			1.00/1.00/1.00	-	0.00/0.00/0.00
NMR$_{n3}$(δ_{C})	1.00/1.00/1.00	0.00/0.00/0.00	0.41/0.56/0.56	0.59/0.44/0.44	0.00/0.00/0.00
			0.97/1.00/1.00	-	0.03/0.00/0.00
NMR(δ_{N})	0.29/028/0.28	0.71/0.72/0.72	0.00/0.00/0.00	0.28/0.15/0.15	0.72/0.85/0.85
			0.08/0.05/0.05	-	0.92/0.95/0.95
NMR$_{n3}$(δ_{N})	0.29/0.28/0.28	0.71/0.72/0.72	0.00/0.00/0.00	0.34/0.20/0.19	0.66/0.80/0.81
			0.10/0.06/0.06	-	0.90/0.94/0.94
NMR(δ_{HC})	1.00/0.93/0.71	0.00/0.07/0.29	0.47/0.85/0.74	0.53/0.10/0.01	0.00/0.05/0.25
			1.00/0.87/0.74	-	0.00/0.13/0.26
NMR$_{n3}$(δ_{HC})	1.00/1.00/0.80	0.00/0.00/0.20	0.41/0.71/0.74	0.59/0.29/0.19	0.00/0.00/0.08
			0.97/0.84/0.75	-	0.03/0.16/0.25
NMR(δ_{HN})	0.29/0.25/0.27	0.71/0.75/0.73	0.00/0.00/0.00	0.28/0.07/0.10	0.72/0.93/0.90
			0.08/0.03/0.04	-	0.92/0.97/0.96
NMR$_{n3}$(δ_{HN})	0.29/0.26/0.27	0.71/0.74/0.72	0.00/0.05/0.04	0.34/0.00/0.03	0.66/0.95/0.93
			0.10/0.05/0.05	-	0.90/0.95/0.95
NMR(δ_{CN})	0.32/0.35/0.31	0.68/0.65/0.69	0.00/0.00/0.00	0.31/0.29/0.20	0.69/0.71/0.80
			0.11/0.13/0.08	-	0.89/0.87/0.92
NMR$_{n3}$(δ_{CN})	0.32/0.36/0.31	0.68/0.64/0.69	0.00/0.00/0.00	0.40/0.45/0.29	0.60/0.55/0.71
			0.12/0.14/0.09	-	0.88/0.86/0.91

Some of the models give implausible results: the models using only ^{1}H chemical shifts always yield a population of 100 % τ-, respectively α+τ-tautomer, regardless of the EC-RISM or normalization method used, which is in contrast to the energetics calculations, which show a significant π-tautomer fraction and a non-neglegible α+π-tautomer fraction of the monocation. The ^{1}H nucleus alone does not seem to be sensitive to tautomer changes. This is plausible since the shifts of the protons that are most likely to be affected by tautomer changes are the tautomerizing protons, and they cannot be detected due to the fact that proton transfer in aqueous solution is fast in relation to the NMR timescale.

In contrast, using only ^{13}C chemical shifts yields a 100% population of the π-tautomer, which again is in contrast to the energetics calculations and literature. In case of the monocation, the models give a mixture of the α+π- and π+τ-tautomers with the main tautomer being the π+τ-tautomer, by not using a normalization, and the α+π-tautomer, applying a normalization. The models which are neglecting the, according to the energetics and literature, not populated π+τ-tautomer result in nearly 100% of the α+π-tautomer. The ^{13}C chemical shifts are likewise not sensitive for the tautomerization process since they are not involved in the process (when the tautomerization takes place at the nitrogen atoms like for histamine, or at the nitrogen and oxygen atoms as it is the case for nucleobases and nucleotides). This may be different if carbon atoms are directly involved in the prototropic tautomerism. In contrast to the ^{1}H models, the wrong main tautomer is predicted, therefore these models should not be used to elucidate tautomer fractions.

The combination of ^{1}H and ^{13}C chemical shifts in a model results in, like for the single models, an insufficient description of the histamine tautomerism. The π-tautomer is the main tautomer of the free base, fully populated without normalization (and with the norm$_{N}$ using EC-RISM/n3) with a rising population of the τ-tautomer upon normalization and going from norm$_{N}$ to norm$_{a}$. For the monocation, like in the ^{13}C only model, a mixture of α+π- and π+τ-tautomers is obtained, with a smaller π+τ-fraction when using the GAFF force field and a normalization, especially norm$_{a}$. The models neglecting the π+τ-tautomer are the only ones with a significant α+τ-fraction; however, it is always, in contrast to the energetics and literature, a minor tautomer.

All of the other models include the ^{15}N chemical shifts. These models are using the ^{15}N chemical shifts only, the combination of these shifts with either the ^{1}H or ^{13}C chemical shifts, or all three types of chemical shifts. All of these models perform similarly: for the free base, the τ-tautomer is identified as the main tautomer together with a significant π-fraction of 25-36%, in which the differences between EC-RISM and EC-

RISM/n3 are negligible. Models with the norm$_N$ normalization yield slightly higher, the ones using norm$_a$ slightly lower π-fractions compared to the use of unnormalized shifts. For the monocation, the α+τ-tautomer is identified as the main tautomer, but no significant α+π-, therefore π+τ-fractions occur, especially when no normalization and EC-RISM/n3 is used within the model. Neglecting this tautomer, the α+π-tautomer is correctly identified as the minor tautomer of the monocation with fractions very close to the computed ones.

This leads to the conclusion that the ^{15}N nucleus is the most sensitive for changes in the NMR chemical shifts upon tautomerization of histamine, and that the chemical shifts of this nucleus always have to be included when building a model to determine the tautomer fractions by fitting to experimental chemical shifts. It is possible to build reliable models using only the ^{15}N chemical shifts, but also combination of these with ^{1}H and/or ^{13}C chemical shifts is possible. The use of a combination with only the ^{13}C chemical shifts is not recommended as the ^{1}H chemical shifts seem to be helpful since only the models using all chemical shifts or the ^{1}H and ^{15}N are able to correctly identify the α+π-tautomer as the minor tautomer in the models, consisting of three tautomers of the histamine monocation. Therefore, a normalization is needed, and the norm$_N$ normalization seems to be the most promising one as it minimizes the π+τ-fractions in the most successful models.

The energetic calculations and literature data as well as the most promising models to extract the tautomer fractions of histamine, the models using the ^{15}N chemical shifts in combination with the ^{1}H and ^{1}H/^{13}C chemical shifts and using the norm$_N$ normalization, are consistent in the τ-tautomer being the main tautomer of the histamine free base together with a minor but significant π-fraction and the α+τ-tautomer being the main tautomer of the monocation, with a small α+π- and no significant π+τ-fraction. The reparametrization of the n3 GAFF atomtype helps to improve the quality of pK_a-calculations for histamine but the influence of the relative energetics of the ionization states is larger than the influence on the tautomer fractions calculated within an ionization state or the NMR chemical shifts. The inclusion of CCSD(T) gas-phase free energies helps to distinguish between the main conformations but has only little impact on the calculated tautomer fractions. It is important to consider the small energetic differences associated with large shifts in tautomer fractions (Table 11), which makes the investigation using multiple methods important. The correlation between experimental NMR chemical shifts of the ionization states and the calculated ones for the main conformations underline the validity of the conformational and tautomeric fractions calculated.

The histamine investigation showed how tautomeric fractions can be calculated using a combined computational and experimental approach and clearly demonstrated that the ^{15}N nucleus is the most important one for the determination of tautomer fractions of nitrogen heterocycles. This is important for the determination of the tautomerism of nucleic acid building blocks and could also help to clarify the problems which occurred with the SAMPL2 dataset. To make these investigations possible, especially at non-ambient conditions, the calculation of NMR chemical shifts has to be possible at these conditions. Therefore, the shielding constants of suitable reference substances have to be calculated, and computational referencing methods have to be investigated in detail. This is done in the next chapter.

4.4 Calibrating the prediction of NMR chemical shifts

4.4.1 High hydrostatic pressure conditions

Calculations at the MP2/6-311+G(d,p) level of theory, introduced during the SAMPL6 challenge,[4] as well as the improved electrostatics model and use of the high pressure PMV correction[7] not only influence the calculation of the energetics of a molecule heavily, they also impact the calculation of ambient pressure and *p*-dependent NMR chemical shifts. Therefore, a recalculation of NMR reference substances is needed. It is not only possible to incorporate pressure and/or temperature effects in the solvent susceptibility used for EC-RISM calculations by using the HNC approximation, but it is also possible to calculate solvent susceptibilities using bridge functions obtained from FF MD simulations. This is done in Ref. 279 and 234 but not used during this work, since the use of the HNC approximation seems to be more reliable. In Ref. 89, the problems which occur by reproducing NMR experiments are outlined. Using a flexible molecule with multiple rotatable bonds, like DSS, the conformational ensemble and its possible modulation by pressure variation have to be considered. Also, the experimental conditions and the referencing (an internal, pressure-dependent, or external, pressure-independent, referencing is possible) used to study the respective nuclei have to be optimally reproduced by the computational workflow. The performance of the new computational framework is studied by investigating the small osmolyte trimethylamine N-oxide (TMAO) and the N-methyl-acetamide (NMA), which is a bioorganic building block used here to mimic the peptide backbone with the smallest molecule possible for computational efficiency.

Addressing the first problem, a population analysis of DSS conformations for varying pressures has been done. The populations and structures are depicted in Figure 21 and show, similar to Ref 89, a clearly favored main conformer. This main conformer has a varying fraction ranging from 0.78 at ambient conditions to 0.69 at 10 kbar. Also four minor populated conformations with fractions of approx. 0.05 (1 bar) increasing to approx. 0.07 (10 kbar) and two nearly unpopulated conformations (fraction of 0.005 (1 bar) to 0.01 (10 kbar)) are observed.[7] The fraction of the main conformer is slightly and steadily decreasing upon pressurization while the other conformers show the opposite trend, their fraction increases. The DSS structures and all raw and processed data are given in the DSS subfolder of SI part 04.

In Figure 21 and Table 14, the pressure dependent shielding constants for the ^{1}H and ^{13}C nuclei of the DSS anion and the ^{15}N nucleus of ammonia are shown. It is important to notice that for the pressure and temperature dependent ^{15}N reference shielding constants, the temperature and pressure dependence of the secondary standard (the -19.4 ppm offset between ammonia in water and liquid ammonia)[295] could play a role; since the pressure and temperature dependence on the shielding constants is very small, this possible influence is neglected here. These DSS shielding constants are the arithmetic mean of the population-weighted shielding constants of the equivalent nuclei from the DSS methyl groups. For the ^{13}C and ^{15}N nuclei, the pressure-dependent shielding constants from EC-RISM calculations, scaled by the pressure dependence of the DSS anion ^{1}H shielding constant, are shown. These shielding constants are only marginally affected by pressure: for the hydrogen atom, an increase from 31.900 (1 bar) to 31.906 ppm (10 kbar) can be observed; the direct shielding of the carbon atom increases from 200.93 (1 bar) to 201.11 ppm (10 kbar), while the indirect, scaled carbon shielding only increases to 200.97 ppm due to the small pressure dependence of the hydrogen nucleus. This small inherent pressure dependence underlines that DSS is an ideal standard for high-pressure NMR experiments. It is a reference substance that allows for reliably probing the pressure-induced changes in the local magnetic field of the solute under study, especially when the indirect referencing via scaling of the ^{13}C resonance by the ^{1}H pressure dependence is used.

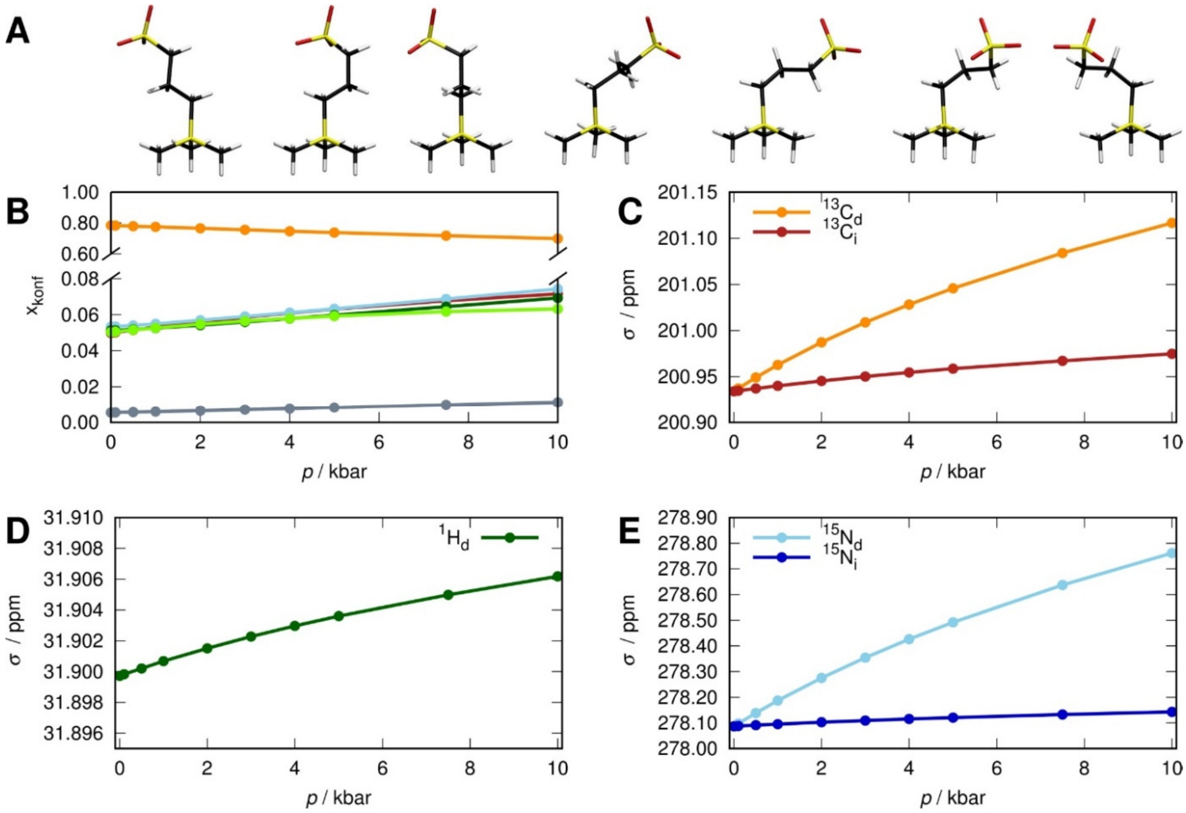


Figure 21: Results of the DSS anion and ammonia shielding calculations. The main conformers of the DSS anion in a descending order of populations (A) and the corresponding pressure-dependent populations of these seven conformers (B) are shown. In addition, the averaged and population-weighted calculated pressure dependent shielding constants of the DSS methyl-group ^{13}C nuclei (C), the DSS methyl-group ^1H nuclei (D), and the ammonia ^{15}N nucleus (E) are depicted. The ^{13}C and ^{15}N shielding constants result from direct calculation (subscript d) and scaling by the pressure dependence of the ^1H DSS shielding (subscript i). Published by Elsevier in Ref 7 (https://www.sciencedirect.com/science/article/abs/pii/S0301462219303412).

In case of ammonia (all raw data are presented in the NH3 subfolder of SI part 04), the pressure dependent increase ranges from 278.09 ppm to 278.76 at 10 kbar for the direct observation of the nucleus, and up to 278.14 ppm while scaling with the ^1H DSS pressure dependence. These observations are made for the ^{15}N nucleus of ammonia in water, while the reference substance for ^{15}N recommended by the IUPAC is liquid ammonia. Therefore, the experimental chemical shift from the ^{15}N nucleus of ammonia in water (-19.4 ppm at 1bar[296]) to liquid ammonia has to be included in this data, which can afterwards directly be used for NMR calculations in water in order to match ^{15}N experiments where direct referencing to liquid ammonia is used, assuming pressure independence of the chemical shift of liquid ammonia to ammonia in water. The introduction of not only an improved computational approach but also the pressure dependent ^{15}N reference shielding constants should allow for the calculation of pressure dependent ^{15}N chemical shifts with increased accuracy compared to prior work.[89]

Table 14: Pressure dependent shielding constants of the reference substances DSS (^{1}H, ^{13}C) and ammonia (^{15}N). The reference shielding constants are given pressure dependent from EC-RISM calculations (subscript d) and scaled by the pressure dependence of the ^{1}H shielding constant (subscript i).

Pressure / kbar	$^{13}C_d$	$^{1}H_d$	$^{15}N_d$	$^{13}C_i$	$^{15}N_i$
0.001	200.934	31.8997	278.087	200.934	278.087
0.1	200.937	31.8998	278.098	200.938	278.098
0.5	200.949	31.9002	278.139	200.952	278.143
1	200.963	31.9007	278.187	200.969	278.196
2	200.987	31.9015	278.275	200.998	278.291
3	201.008	31.9023	278.354	201.024	278.377
4	201.027	31.9030	278.426	201.048	278.455
5	201.045	31.9036	278.492	201.070	278.526
7.5	201.082	31.9050	278.637	201.115	278.684
10	201.115	31.9062	278.762	201.156	278.819

This data allows a further methodological extension related to the indirect experimental reference method and how it can be optimally mapped to a computational framework. In the experimental practice, an Ξ factor is used to scale the resonance of a nucleus other than hydrogen by the DSS ^{1}H resonance. Here, three different approaches are used for referencing pressure-dependent NMR experiments: The first one is the use of the 1 bar resonance by referencing to the target nucleus (^{13}C or ^{15}N), which implies a pressure independent Ξ factor and will be termed "σ_{ref}(1 bar)". Second, a scaling of the reference nucleus shielding constant at 1 bar by $\sigma(^{1}\text{H}, p)/\sigma(^{1}\text{H}, 1\text{ bar})$ is used here, which should more closely match the experimental setup; this variant will be denoted as "$\sigma_{\text{ref,i}}$(p)". The third method is the direct referencing to a pressure dependent standard and will be called "$\sigma_{\text{ref,d}}$(p)".

For both of the benchmark molecules TMAO and NMA, experimental reference data are available in Ref. 7 (TMAO) and Ref. 234 (NMA). In Figure 22, the pressure dependence of the chemical shifts and of the shielding constants of the TMAO ^{1}H, ^{13}C, and ^{15}N nuclei are shown. The structures and full chemical shift data are given in the TMAO subfolder of SI part 04. These chemical shifts are, in agreement with the experiment, only slightly pressure dependent. The absolute shifts are in the same order of magnitude as the experiment, but the trends of the pressure dependence have to be investigated. Experimentally, the chemical shifts increase with increasing pressure for all nuclei studied, which is only computationally reflected correctly by using direct referencing with a pressure dependent standard. This is also underlined by the matching sign of the linear and quadratic pressure coefficients from the polynomial fits shown in Table 15. Unique is the ^{13}C nucleus; it is the only nucleus where only the direct referencing variant gives the correct sign of the slope, also here the shielding constant is found to increase with rising pressure in contrast to the other nuclei. This clearly shows the need for a pressure dependent computational referencing method to match the sign of the slope, even though this contradicts the experimental referencing practice. Generally, the direct referencing in the calculation always results in larger slopes compared to the indirect and the pressure independent variants and in an overall better agreement with the experiments. Thereby, only the ^{1}H chemical shift slopes are strongly overestimated, which is unexpected because the protons are directly referenced in the experiments.

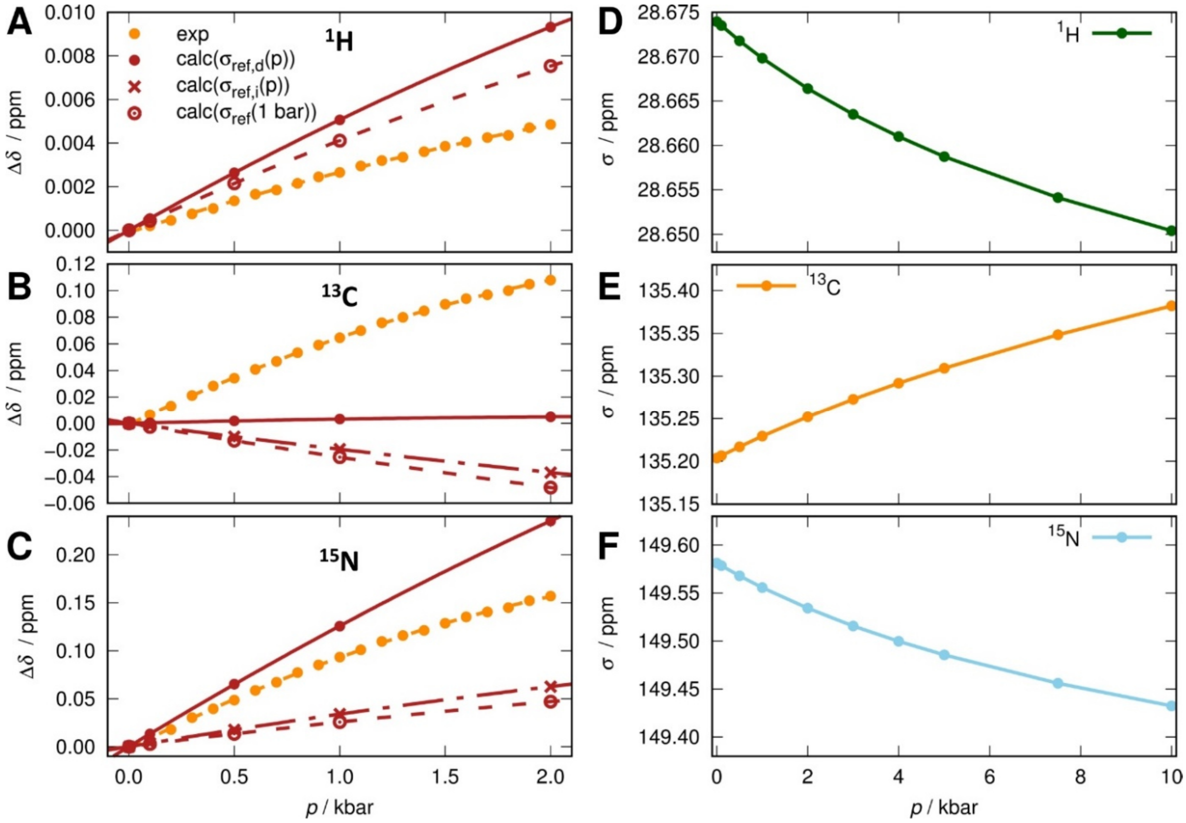


Figure 22: Pressure dependence of experimental (yellow) and calculated (red) chemical shifts of the ^{1}H (A), ^{13}C (B) and ^{15}N (C) nuclei of TMAO, relative to 1 bar. The DSS and ammonia shielding constants shown in Figure 21 are used for the referencing, which is done pressure dependent (solid), with respect to the 1 bar reference shielding constants (dashed), and using the 1 bar reference shieldings scaled by the ^{1}H pressure dependence (dashed-dotted). In addition, the second order polynomial fits (see Table 15) and the corresponding shielding constants of the TMAO ^{1}H (D), ^{13}C (E) and ^{15}N (F) nuclei are shown. Published by Elsevier in Ref 7 (https://www.sciencedirect.com/science/article/abs/pii/S0301462219303412).

In Figure 24, the corresponding data for NMA is shown. NMA has two different conformations, the *cis*- and the *trans* conformation, which is the main conformation and the smallest protein backbone mimic; both conformations are depicted in Figure 23.

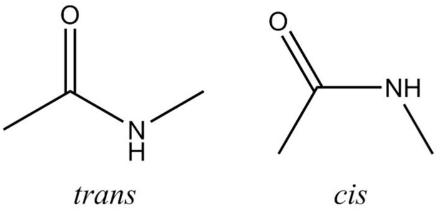


Figure 23: Conformations of N-methyl-acetamide, the molecule is the smallest unit of a protein backbone. The trans conformation is the dominant species.

For both conformations, the experimentalists were able to identify the respective signals, so no calculation of population-weighted shielding constants and shifts is needed here, and it is possible to judge the quality of NMR calculations independent of the calculated energetics. The ambient condition chemical shifts of the ^{1}H and ^{15}N nucleus are higher and the one of the ^{13}C is lower for the *trans* compared to the *cis* conformation; the trend for the shielding constants is opposite. This is in contrast to the experimental findings, but the experimental and calculated shifts are in the same order of magnitude, and to judge the different referencing methods, the pressure dependence has to be considered.

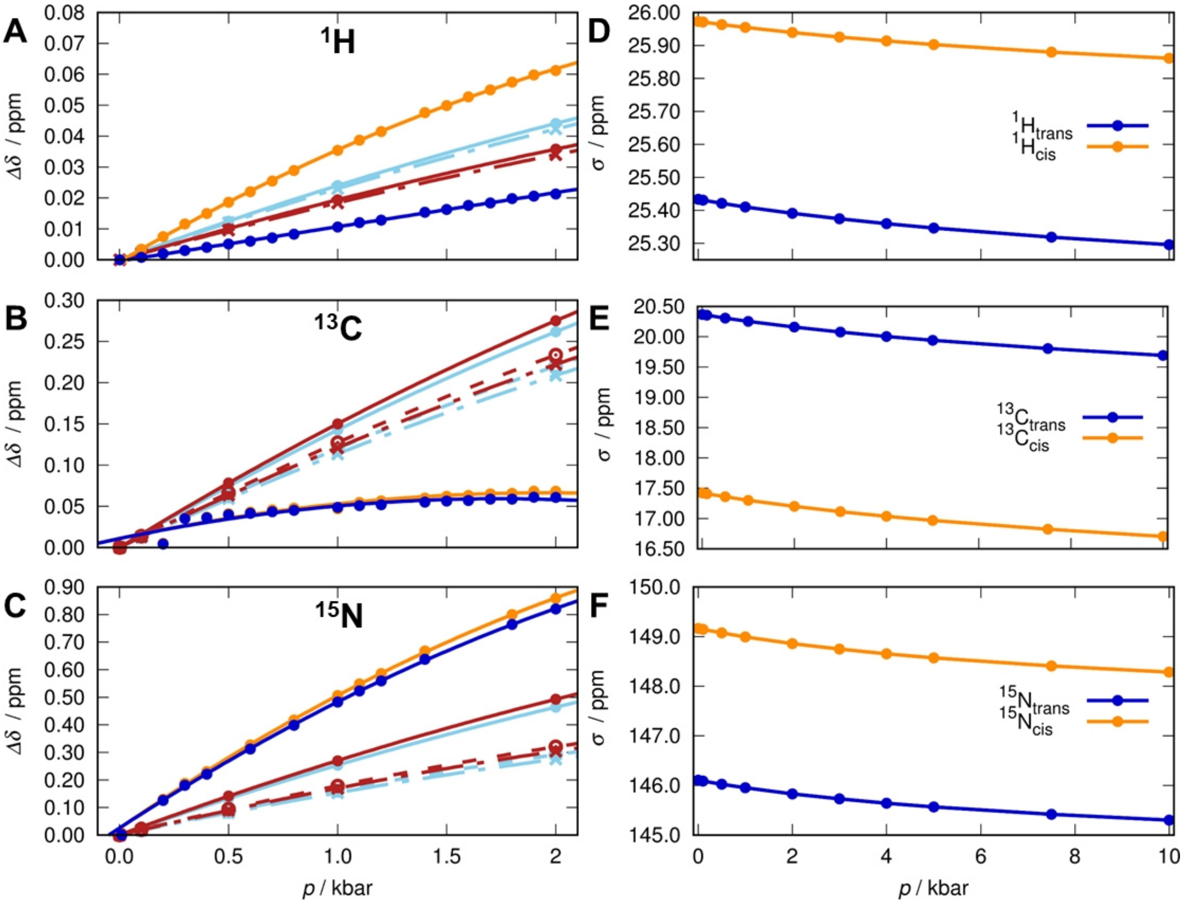


Figure 24: Pressure dependence of experimental and calculated cis (yellow and red) and trans NMA (blue and light blue) chemical shifts of the ^{1}H (A), ^{13}C (B) and ^{15}N (C) nuclei relative to 1 bar. The DSS and ammonia shielding constants shown in Figure 21 are used for the referencing, which is done pressure dependent (solid), with respect to the 1 bar reference shielding constants (dashed), and using the 1 bar reference shieldings scaled by the ^{1}H pressure dependence (dashed-dotted). In addition, the second order polynomial fits (see Table 15) and the corresponding shielding constants of the NMA ^{1}H (D), ^{13}C (E) and ^{15}N (F) nuclei are shown.

In contrast to TMAO, the chemical shifts of all nuclei increase upon pressurization, this is consistent between experiment and calculation. These pressure dependent trends are small and monotonous. For the ^{1}H nucleus, the pressure dependence is slightly overestimated for the *trans* conformation and underestimated for the *cis* conformation. Therefore, by resulting in larger chemical shifts, the direct pressure dependent referencing reflects the experimental findings slightly better for the *cis* than for the *trans* conformation, but the differences are very small. In case of the ^{13}C nucleus, the pressure dependence is computationally overestimated. This effect results in the smallest calculated shifts, using the pressure independent referencing, having the smallest deviation from the experiment. The pressure dependence of the ^{15}N chemical shifts is slightly underestimated by the computation. Here again, the pressure dependent referencing shows the best agreement with the experiment, underlining the usefulness of the newly developed pressure dependent nitrogen referencing.

Table 15: Linear (B_1, in ppm kbar^{-1}) and quadratic (B_2, in ppm kbar^{-2}) coefficients from fitting experimental and calculated TMAO and NMA amide chemical shifts to the following form: $\delta(p)=\delta_0+B_1p+B_2p^2$. In addition, the 1 bar chemical shifts (δ_0 in ppm) are shown.

	B_1(^{1}H)	B_2(^{1}H)	B_1(^{13}C)	B_2(^{13}C)	B_1(^{15}N)	B_2(^{15}N)
TMAO/Exp.	0.00285	-0.00021	0.07501	-0.01047	0.10682	-0.01415
TMAO/$\sigma_{\mathrm{ref,d}}$(p)	0.00547	-0.00040	0.00389	-0.00071	0.13398	-0.00817
TMAO/$\sigma_{\mathrm{ref,i}}$(p)	0.00547	-0.00040	-0.02050	0.00101	0.03651	-0.00161
TMAO/σ_{ref}(1 bar)	0.00447	-0.00035	-0.02680	0.00134	0.02779	-0.00214
NMA(*trans*)/Exp.	0.01093	0.00005	0.06960	-0.02152	0.54669	-0.07123
NMA(*trans*)/$\sigma_{\mathrm{ref,d}}$(p)	0.02597	-0.00196	0.15447	-0.01178	0.27554	-0.02180
NMA(*trans*)/$\sigma_{\mathrm{ref,i}}$(p)	0.02597	-0.00196	0.12980	-0.00986	0.17870	-0.01653
NMA(*trans*)/σ_{ref}(1 bar)	0.02488	-0.00187	0.12294	-0.00929	0.16921	-0.01573
NMA(*cis*)/Exp.	0.04099	-0.00490	0.07075	-0.02007	0.57450	-0.07545
NMA(*cis*)/$\sigma_{\mathrm{ref,d}}$(p)	0.02093	-0.00151	0.16261	-0.01251	0.29368	-0.02370
NMA(*cis*)/$\sigma_{\mathrm{ref,i}}$(p)	0.02093	-0.00151	0.13794	-0.01060	0.19683	-0.01842
NMA(*cis*)/σ_{ref}(1 bar)	0.01984	-0.00142	0.13108	-0.01002	0.18735	-0.01762
	^{1}H		^{13}C		^{15}N	
TMAO/δ_0 (exp)	3.250		62.251		104.572	
TMAO/δ_0 (calc)	3.226		65.730		109.106	
NMA(*trans*)/δ_0 (exp)	7.097		179.999		111.949	
NMA(*trans*)/δ_0 (calc)	6.466		180.566		112.581	
NMA(*cis*)/δ_0 (exp)	7.843		177.298		113.763	
NMA(*cis*)/δ_0 (calc)	5.927		183.512		109.524	

With the exception of ^{13}C, the NMA chemical shifts are more sensitive to pressure changes than the ones of

TMAO and, omitting the experimental second-order ^{1}H pressure coefficient of the trans conformation (which is special by having a much smaller absolute value than the *cis*-conformation and the computations), all signs and orders of magnitude of the experimental pressure-coefficients are reproduced by all three computational approaches.

Overall, it can be observed that direct referencing in the computations always results in larger slopes and a stronger pressure dependence of the chemical shifts compared to the indirect and especially the pressure independent variants. This indicates that the slight pressure dependence of the reference substance has to be taken into account. Summarizing, using pressure dependent standards in the calculations is apparently a reliable strategy if one wants to study various nuclei on the same theoretical footing, which is now also possible for ^{15}N, and the absolute numbers are closer to the experiment compared to previous work.[7,89] Other referencing approaches, using explicit solvation, are investigated in detail in S. Maste's master's thesis.[297]

4.4.2 Variation of temperature conditions

Like for referencing the high pressure conditions, a similar approach has to be used for referencing NMR shielding constants at different temperatures. Here, some additional problems have to be faced: the need for a temperature dependent PMV correction, and the need for temperature dependent TMAO and NMA measurements. The temperature dependent PMV correction is developed in this work and described in chapter 4.1 used here for the calculation of the temperature dependent DSS and NMA populations. This order was chosen because the NMR calculations are done in a complementary way to the high pressure calculations. Due to the fact that no experimental data for the temperature dependence of NMA and TMAO chemical shifts is available, a comparison with this data to judge the different referencing methods is not possible, but the trends that have to be expected from further calculations and experiments can be investigated.

The population analysis of DSS conformations for temperature variations and the main conformations are shown in Figure 25, and the results are quite similar to those for the high pressure calculations. There is a clearly favored main conformer, and the ranking of conformations stays the same. This main conformer has a fraction of 0.83 at 278.15 K to 0.67 at 372.756 K and, thus, shows a slightly stronger temperature than pressure dependence. The four minor populated conformations have fractions of about 0.040 (278.15 K), increasing to approx. 0.078 (372.756 K). The two nearly unpopulated conformations have fractions of

0.0034 (278.15 K) to 0.0133 (372.756 K). Again, the fraction of the main conformer is slightly and steadily decreasing with increasing temperature, and the other conformers behave in the opposite way. Generally, the temperature dependence of conformational fractions is stronger than the pressure dependence, even though the temperature range covers approx. 100 K in contrast to the investigated pressure range of 10 kbar. Now, the comparison of the tautomer populations at ambient conditions, calculated with the slightly different solvent susceptibilities of the pressure and temperature series can be done. Using the pressure dependent solvent susceptibilities, the DSS main conformation has a fraction of 0.78. This is in good agreement with the results obtained from the temperature dependent calculations with 0.79; the energy differences resulting in this change of population are negligible, also the energetics and the energetic ordering of the minor tautomers is nearly unaffected by the switch of solvent susceptibilities. The population-weighted shielding constants of the ^{1}H and ^{13}C nuclei of the DSS anion and the ^{15}N nucleus of ammonia are depicted in Figure 25. For the ^{13}C and ^{15}N nuclei, the direct EC-RISM shielding constants and the scaled ones, this time by the temperature dependence of the DSS anion ^{1}H shielding constant, are shown. The temperature dependence of these shielding constants is in a similar range as the pressure dependence. The hydrogen shielding shows an increase from 31.899 (278.15 K) to 31.902 ppm (372.756 K), the shielding of the carbon atom decreases from 200.97 (278.15 K) to 200.80 ppm (372.756 K) for the directly calculated one and to 200.81 ppm for the scaled one. By having an inherent temperature dependence in the same range as the pressure dependence, the DSS is also well suited as a reference substance for temperature variations in NMR experiments.

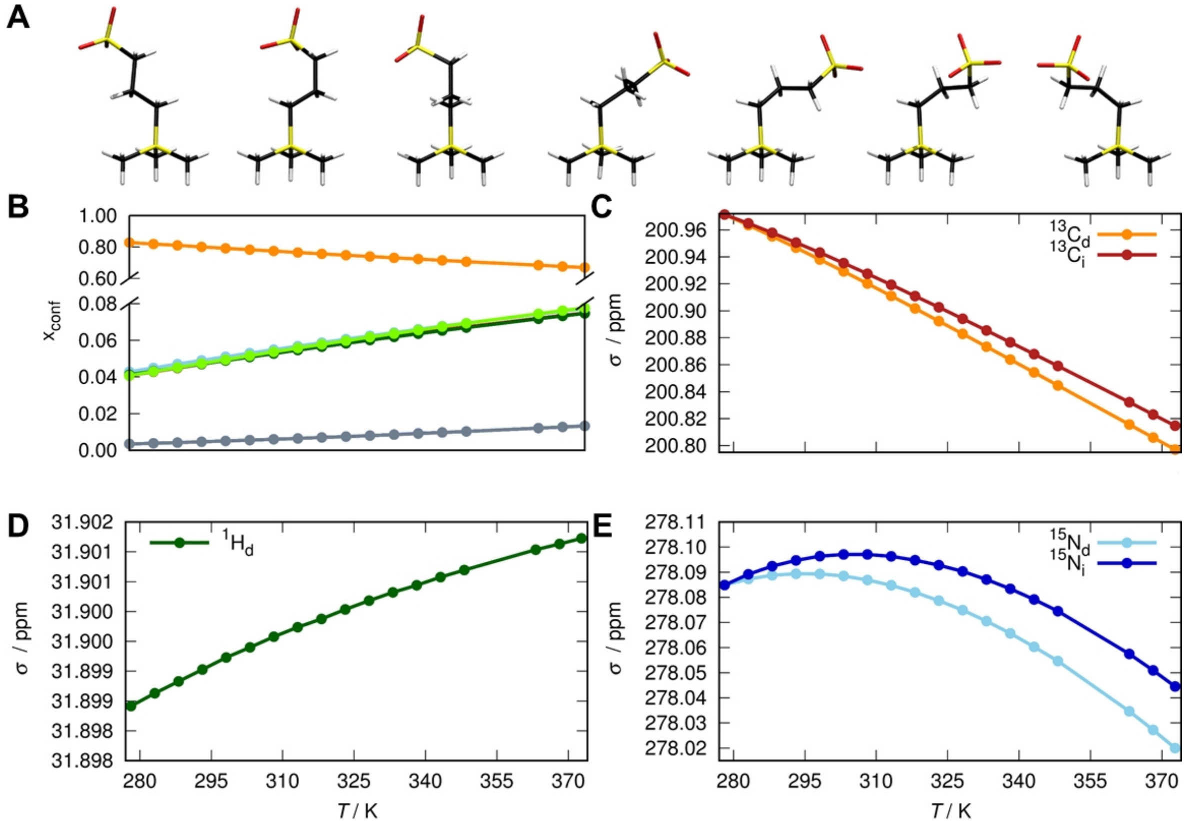


Figure 25: Results of the DSS anion and ammonia shielding calculations. The main conformers of the DSS anion in a descending order of populations (A) and the corresponding temperature-dependent populations of this seven conformers (B) are shown. In addition, the averaged and population-weighted calculated temperature-dependent shielding constants of the DSS methyl-group ^{13}C nuclei (C), the DSS methyl-group ^{1}H nuclei (D), and the ammonia ^{15}N nucleus (E) are depicted. The ^{13}C and ^{15}N shielding constants result from direct calculation (subscript d) and scaled by the temperature dependence of the ^{1}H DSS shielding (subscript i).

The ammonia shielding, too, shows a very small temperature dependence but does not behave in a linear way; the shielding constant increases from 278.085 ppm at 278.15 K to 278.01 ppm at 308.15 K and afterwards decreases down to 278.02 ppm for the direct calculation and down to 278.055 ppm by scaling with the ^{1}H temperature dependence at 372.756 K. This non-linear behavior of temperature dependent ^{15}N chemical shifts is not found quantitatively in the literature, but the overall trend, decrease in the shielding constant and therefore an increase in the chemical shift with increasing temperature, is consistent with the trends presented in the literature (Ref. 295 and 298). A possible reason for the non-linear behavior may be the temperature dependence of the chemical shift difference of the ^{15}N nucleus of ammonia in water (-19.4 ppm at 298 K[296]) to liquid ammonia, which is, like for the pressure dependence, assumed to be temperature independent in this work.

Table 16: Temperature dependent shielding constants of the reference substances DSS (^{1}H, ^{13}C) and ammonia (^{15}N). The reference shielding constants are given temperature dependent from EC-RISM calculations (subscript d) and scaled by the temperature dependence of the ^{1}H shielding constant (subscript i).

Temperature / K	$^{13}C_d$	$^{1}H_d$	$^{15}N_d$	$^{13}C_i$	$^{15}N_i$
278.15	200.971	31.8989	278.085	200.971	278.085
283.15	200.963	31.8991	278.087	200.965	278.089
288.15	200.955	31.8993	278.089	200.958	278.092
293.15	200.947	31.8995	278.089	200.951	278.095
298.15	200.938	31.8997	278.089	200.943	278.096
303.15	200.929	31.8999	278.089	200.935	278.097
308.15	200.920	31.9001	278.087	200.927	278.097
313.15	200.911	31.9002	278.085	200.919	278.096
318.15	200.902	31.9004	278.082	200.911	278.095
323.15	200.892	31.9005	278.079	200.903	278.093
328.15	200.883	31.9007	278.075	200.894	278.090
333.15	200.873	31.9008	278.071	200.885	278.087
338.15	200.864	31.9009	278.066	200.877	278.083
343.15	200.854	31.9011	278.060	200.868	278.079
348.15	200.845	31.9012	278.055	200.859	278.074
363.15	200.816	31.9015	278.035	200.832	278.057
368.15	200.806	31.9016	278.027	200.823	278.051
372.75	200.797	31.9017	278.020	200.815	278.044

The small inherent temperature dependence of the reference shielding constants leads to the assumption that the choice of the referencing method (the direct temperature dependent referencing, scaling by the temperature dependence of the ^{1}H nucleus and using the 278.15 K shielding) affects the calculated shifts only slightly.

The chemical shifts of TMAO are shown in Figure 26. The ^{1}H chemical shift is decreasing with increasing temperature in contrast to an increase upon pressurization, and the signal is more sensitive to the temperature changes than to pressure changes. The difference between both referencing methods is small with a slightly smaller temperature dependence using the direct referencing. The ^{13}C shielding constant is, also in contrast to the pressure dependent calculations, decreasing with temperature. This should result in an increase of the chemical shift, which is only reflected by using the indirect temperature independent referencing. Using the direct or indirect temperature dependent referencing, the chemical shift is slightly decreasing. Overall, the temperature dependence of the ^{13}C shielding is in the same range as the pressure dependence, but a dependence of the chemical shift slopes on the referencing method can be observed, which is not expected from the reference substance calculations. The findings for the ^{15}N nucleus are in contrast to those for the other nuclei, since the shielding constant is less temperature than pressure dependent. Upon pressurization, the shielding is steadily declining while the temperature dependence is, like for ammonia, first slightly increasing and afterwards decreasing. This leads to an expectation of a first decreasing and afterwards increasing chemical shift. This is, again, only reflected by the temperature independent referencing. The shielding is nearly unaffected by temperature, resulting in a dominance of the ammonia shielding during the referencing for the direct and the scaled approach. Experimental ^{15}N chemical shifts usually increase with temperature;[295,298] since this is only reflected by the temperature independent referencing, this method is recommended for investigating ^{15}N chemical shifts. To finally judge the referencing methods, temperature dependent NMR experiments of TMAO (and NMA) would be helpful.

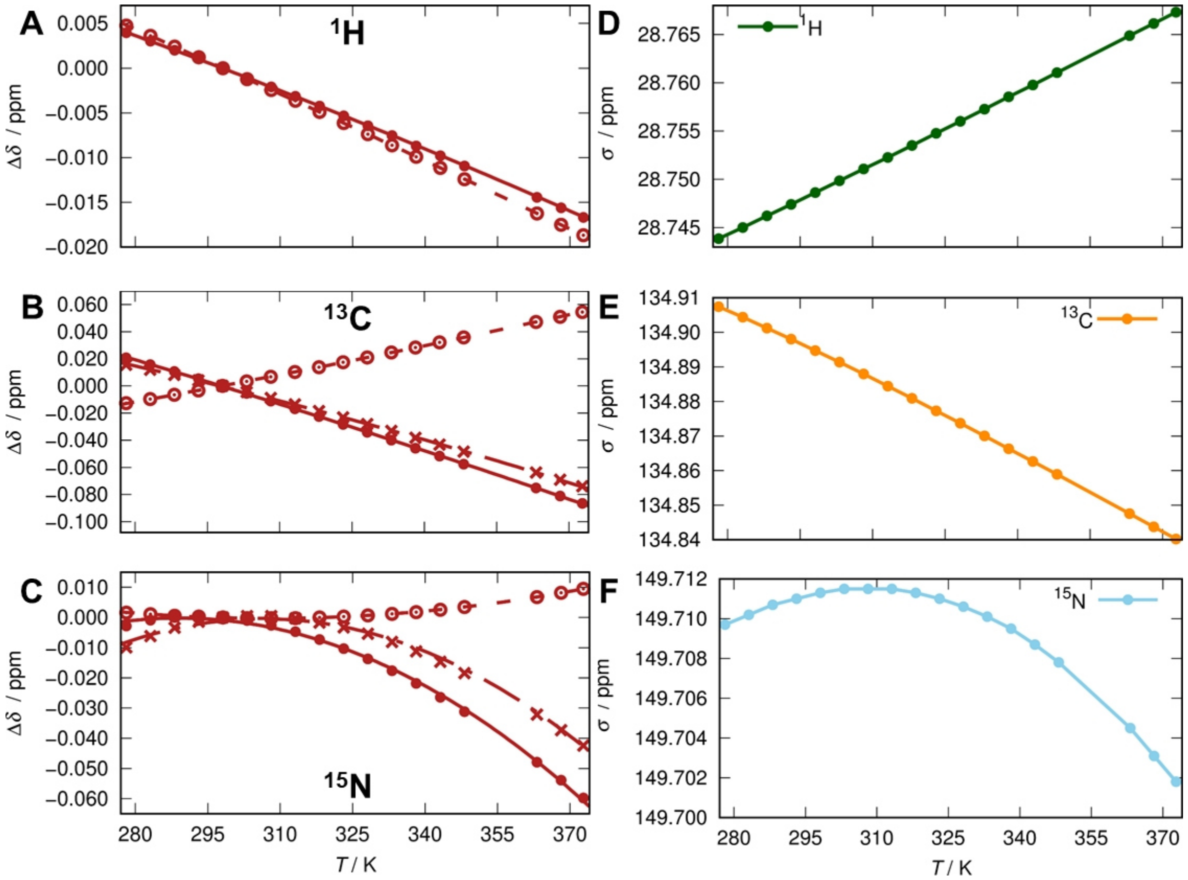


Figure 26: Temperature dependence of calculated (red) chemical shifts of the ^{1}H (A), ^{13}C (B) and ^{15}N (C) nuclei of TMAO, relative to 298.15 K. The DSS and ammonia shielding constants shown in Figure 25 are used for the referencing, which is done temperature dependent (solid line, points), with respect to the 278.15 K reference shielding constants (dashed line, circles), and using the 278.15 K reference shielding constants scaled by the ^{1}H temperature dependence (dashed-dotted line, crosses). In addition, the second order polynomial fits and the corresponding shielding constants of the TMAO ^{1}H (D), ^{13}C (E) and ^{15}N (F) nuclei are shown (see Table 19).

The results of the NMA calculations are shown in Figure 27. The differences in *cis* and *trans* shielding constants are similar to those shown in Figure 24 but all slightly and steadily increasing with temperature. The overall temperature dependence is bigger than in TMAO, and the differences in the referencing methods are, as expected, very small. For all nuclei, the conformation with the higher shielding shows less temperature dependence. In general, the direct referencing results in the strongest temperature dependence, and the 278.15 K referencing in the smallest. Experiments are needed to determine which is in closer agreement to the experimental conditions. But due to the small inherent temperature dependence of the reference substance, the impact of the referencing method is negligible as soon as the molecule investigated exhibits a temperature dependence w.r.t shielding constants. The temperature dependence of the ^{15}N chemical shift of liquid ammonia is reported as 40 ppb,[299] which shows that the temperature dependence of the IUPAC reference substance is also very small. The results are summarized in Table 19.

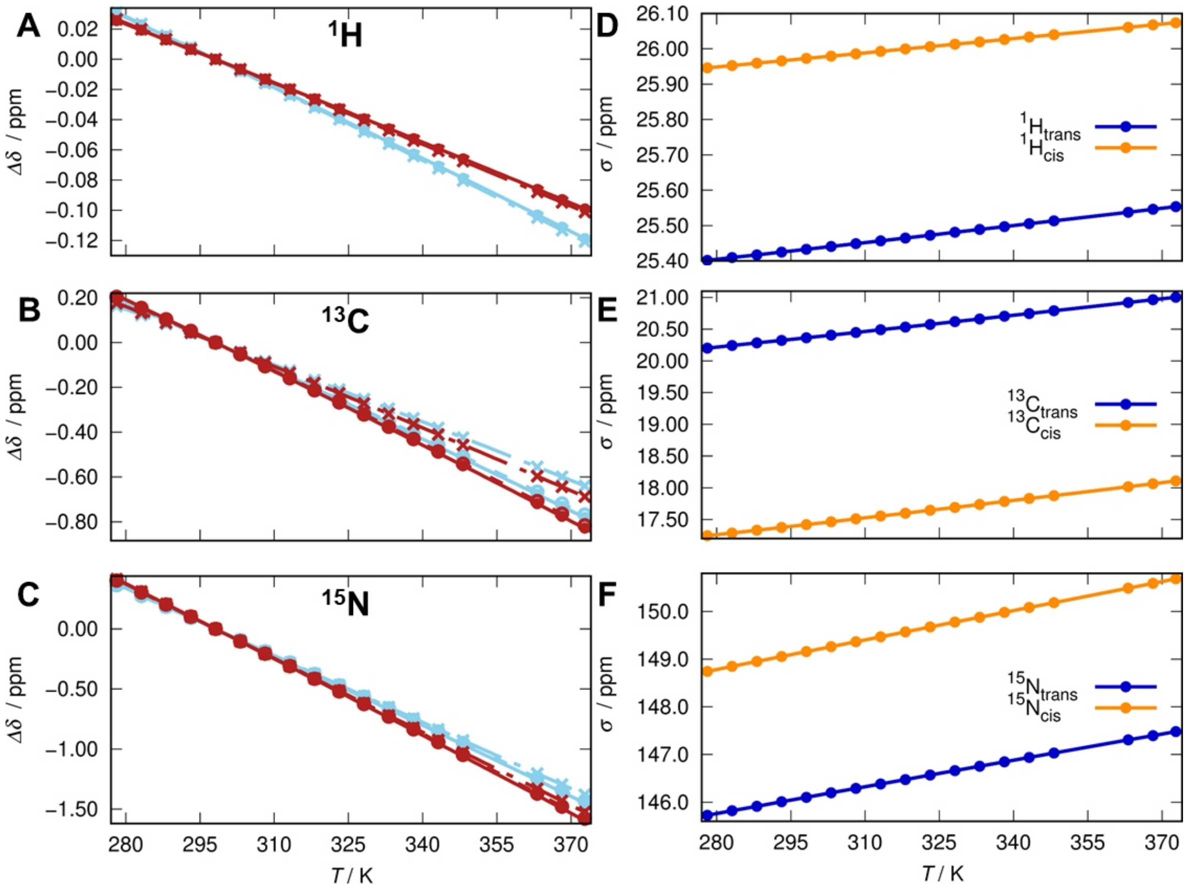


Figure 27: Temperature dependence of calculated NMA cis (yellow and red) and trans (blue and light blue) chemical shifts of the ^{1}H (A), ^{13}C (B) and ^{15}N (C) nuclei relative to 298.15 K. The DSS and ammonia shielding constants shown in Figure 25 are used for the referencing, which is done temperature dependent (solid line, points), with respect to the 278.15 K reference shielding constants (dashed line, circles) and using the 278.15 reference shielding constants scaled by the ^{1}H temperature dependence (dashed-dotted line, crosses). In addition, the second order polynomial fits and the corresponding shielding constants of the NMA ^{1}H (D), ^{13}C (E) and ^{15}N (F) nuclei are shown (see Table 19).

Furthermore, the *cis*/*trans*-NMA populations are interesting due to their dependence on pressure and temperature. This can be seen by an increase in population of the minor conformation, the *cis* conformation (Table 17, Table 18 and Figure 28). Upon pressurization, the *cis*-NMA fraction is increasing from 0.0217 at 1 bar to 0.0329 at 10 kbar while it increases from 0.0163 at 278.15 K to 0.0451 at 372.756 K with rising temperature. The temperature dependence of the populations is larger compared to the pressure dependence. The *trans* conformation mimics the natural peptide bond found in protein backbones, and a significant change in populations, leading to a large *cis* fraction, could be part of the pressure or temperature unfolding processes of proteins. From the results presented here, especially from the temperature dependent calculations, this mechanism may play an important role during the unfolding process. The different solvent susceptibilities used for the calculation of ambient conditions within the respective pressure and temperature series do not change the results; even though the absolute values differ, the energy differences between the conformations remain constant with 2.258 (pressure dependent) and 2.256 kcal/mol (temperature dependent), respectively.

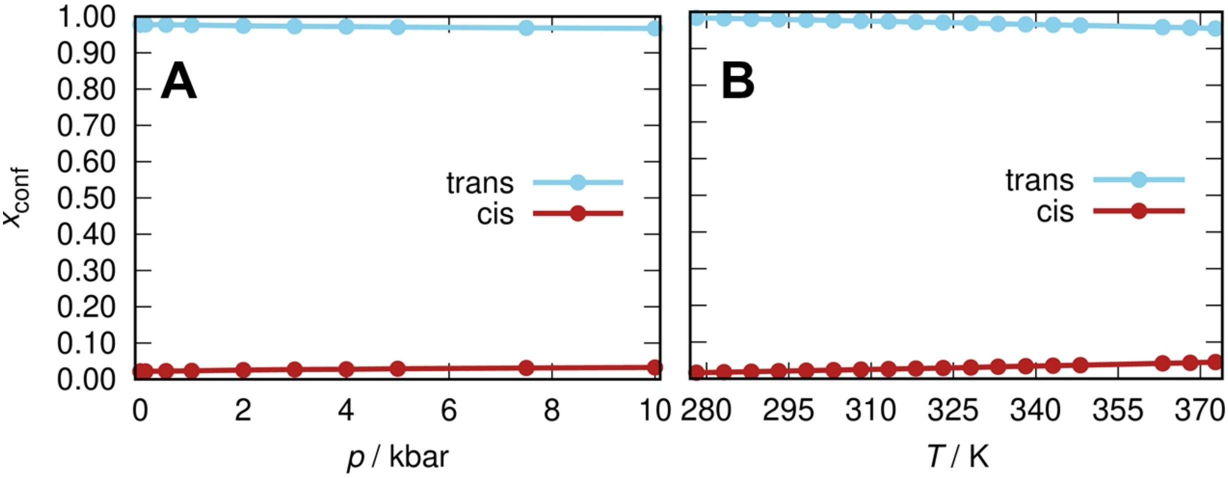


Figure 28: Pressure (A) and temperature (B) dependent populations of cis- and trans NMA. Calculated with EC-RISM using the respective pressure and temperature dependent PMV corrections. The structures and raw data are presented in the NMA subfolder of SI part 04.

Table 17: Energetics of NMA at different pressure conditions. Intramolecular energies in solution (E_{solv} in kcal/mol), excess chemical potentials (μ^{ex} and $\mu^{ex,corr}$ in kcal/mol) with and without PMV correction for the respective conditions as well as the free energy in solution (G_{sol} in kcal/mol) and the conformer fractions calculated using these corrections are given.

	E_{solv}	μ^{ex}	$\mu^{ex,corr}$	G_{sol}	x_{conf}
NMA(*trans,*1 bar)	-155603.298	-12.159	-23.684	-155626.982	0.9783
NMA(*cis,*1 bar)	-155601.175	-12.230	-23.550	-155624.724	0.0217
NMA(*trans,*100 bar)	-155603.275	-11.892	-23.514	-155626.789	0.9781
NMA(*cis,*100 bar)	-155601.153	-11.964	-23.383	-155624.536	0.0219
NMA(*trans,*500 bar)	-155603.188	-10.817	-22.829	-155626.017	0.9773
NMA(*cis,*500 bar)	-155601.072	-10.891	-22.707	-155623.779	0.0227

	E_{solv}	μ^{ex}	$\mu^{ex,corr}$	G_{sol}	x_{conf}
NMA(*trans,*1 kbar)	-155603.088	-9.488	-21.980	-155625.068	0.9763
NMA(*cis,*1 kbar)	-155600.979	-9.565	-21.869	-155622.848	0.0237
NMA(*trans,*2 kbar)	-155602.913	-6.885	-20.322	-155623.234	0.9747
NMA(*cis,*2 kbar)	-155600.816	-6.966	-20.226	-155621.042	0.0253
NMA(*trans,*3 kbar)	-155602.762	-4.359	-18.726	-155621.487	0.9733
NMA(*cis,*3 kbar)	-155600.676	-4.444	-18.642	-155619.318	0.0267
NMA(*trans,*4 kbar)	-155602.629	-1.908	-17.195	-155619.825	0.9721
NMA(*cis,*4 kbar)	-155600.554	-1.996	-17.121	-155617.675	0.0279
NMA(*trans,*5 kbar)	-155602.511	0.471	-15.730	-155618.241	0.9711
NMA(*cis,*5 kbar)	-155600.445	0.380	-15.663	-155616.107	0.0289
NMA(*trans,*7.5 kbar)	-155602.261	6.131	-12.330	-155614.591	0.9689
NMA(*cis,*7.5 kbar)	-155600.215	6.035	-12.275	-155612.490	0.0311
NMA(*trans,*10 kbar)	-155602.057	11.446	-9.245	-155611.303	0.9671
NMA(*cis,*10 kbar)	-155600.027	11.346	-9.199	-155609.226	0.0329

Table 18: Energetics of NMA at different temperature conditions. Intramolecular energies in solution (E_{solv} in kcal/mol), excess chemical potentials (μ^{ex} and $\mu^{ex,corr}$ in kcal/mol) with and without PMV correction for the respective conditions as well as the free energy in solution (G_{sol} in kcal/mol) conformer fractions calculated using these corrections are given.

	E_{solv}	μ^{ex}	$\mu^{ex,corr}$	G_{sol}	x_{conf}
NMA(*trans,*278.15 K)	-155602.936	-13.889	-25.093	-155628.029	0.9837
NMA(*cis,*278.15 K)	-155600.822	-13.942	-24.940	-155625.762	0.0163
NMA(*trans,*283.15 K)	-155603.025	-13.406	-24.716	-155627.741	0.9824
NMA(*cis,*283.15 K)	-155600.909	-13.463	-24.568	-155625.477	0.0176
NMA(*trans,*288.15 K)	-155603.114	-12.947	-24.364	-155627.478	0.9811
NMA(*cis,*288.15 K)	-155600.996	-13.009	-24.222	-155625.217	0.0189
NMA(*trans,*293.15 K)	-155603.204	-12.511	-24.036	-155627.239	0.9797
NMA(*cis,*293.15 K)	-155601.083	-12.578	-23.898	-155624.981	0.0203
NMA(*trans,*298.15 K)	-155603.293	-12.096	-23.730	-155627.023	0.9783
NMA(*cis,*298.15 K)	-155601.170	-12.168	-23.597	-155624.767	0.0217
NMA(*trans,*303.15 K)	-155603.382	-11.701	-23.445	-155626.827	0.9769
NMA(*cis,*303.15 K)	-155601.256	-11.777	-23.316	-155624.572	0.0231
NMA(*trans,*308.15 K)	-155603.471	-11.325	-23.179	-155626.650	0.9754
NMA(*cis,*308.15 K)	-155601.343	-11.405	-23.054	-155624.397	0.0246
NMA(*trans,*313.15 K)	-155603.560	-10.967	-22.932	-155626.492	0.9739
NMA(*cis,*313.15 K)	-155601.430	-11.051	-22.810	-155624.240	0.0261
NMA(*trans,*318.15 K)	-155603.649	-10.625	-22.702	-155626.352	0.9724
NMA(*cis,*318.15 K)	-155601.516	-10.714	-22.584	-155624.100	0.0276
NMA(*trans,*323.15 K)	-155603.738	-10.300	-22.490	-155626.228	0.9709
NMA(*cis,*323.15 K)	-155601.602	-10.392	-22.375	-155623.977	0.0291
NMA(*trans,*328.15 K)	-155603.826	-9.990	-22.294	-155626.121	0.9693
NMA(*cis,*328.15 K)	-155601.687	-10.086	-22.182	-155623.870	0.0307
NMA(*trans,*333.15 K)	-155603.915	-9.694	-22.114	-155626.029	0.9677
NMA(*cis,*333.15 K)	-155601.773	-9.795	-22.004	-155623.777	0.0323
NMA(*trans,*338.15 K)	-155604.002	-9.413	-21.949	-155625.952	0.9661
NMA(*cis,*338.15 K)	-155601.858	-9.518	-21.842	-155623.700	0.0339
NMA(*trans,*343.15 K)	-155604.090	-9.145	-21.799	-155625.889	0.9645

	E_{solv}	μ^{ex}	$\mu^{ex,corr}$	G_{sol}	x_{conf}
NMA(*cis,*343.15 K)	-155601.943	-9.254	-21.693	-155623.636	0.0355
NMA(*trans,*348.15 K)	-155604.178	-8.891	-21.662	-155625.840	0.9629
NMA(*cis,*348.15 K)	-155602.027	-9.004	-21.559	-155623.586	0.0371
NMA(*trans,*363.15 K)	-155604.438	-8.203	-21.334	-155625.773	0.9581
NMA(*cis,*363.15 K)	-155602.279	-8.326	-21.235	-155623.515	0.0419
NMA(*trans,*368.15 K)	-155604.525	-7.997	-21.251	-155625.775	0.9564
NMA(*cis,*368.15 K)	-155602.363	-8.124	-21.153	-155623.516	0.0436
NMA(*trans,*372.756 K)	-155604.604	-7.817	-21.184	-155625.789	0.9549
NMA(*cis,*372.756 K)	-155602.440	-7.948	-21.087	-155623.527	0.0451

Table 19: Intercepts (B_0 in ppm), linear (B_1, in ppm kbar^{-1}) and quadratic (B_2, in ppm kbar^{-2}) coefficients from fitting calculated TMAO and NMA amide chemical shifts to the following form: $\delta(p)=\delta_0+B_0+B_1T+B_2T^2$. Also the 278.15 K chemical shifts (δ_0 in ppm) are shown.

	B_0(^{1}H)	B_1(^{1}H)	B_2(^{1}H)	B_0(^{13}C)	B_1(^{13}C)	B_2(^{13}C)	B_0(^{15}N)	B_1(^{15}N)	B_2(^{15}N)
TMAO/$\sigma_{\mathrm{ref,d}}$(T)	0.03670	-0.00007	$-2.3\cdot10^{-7}$	0.23146	-0.00060	$-8.4\cdot10^{-7}$	-0.71960	0.00500	$-8.6\cdot10^{-6}$
TMAO/$\sigma_{\mathrm{ref,i}}$(T)	0.03670	-0.00007	$-2.3\cdot10^{-7}$	0.09703	-0.00011	$-1.6\cdot10^{-6}$	-0.90555	0.00597	$-9.8\cdot10^{-6}$
TMAO/σ_{ref}(278.15K)	0.05802	-0.00018	$-1.1\cdot10^{-7}$	-0.11434	0.00018	$8.2\cdot10^{-7}$	0.21477	-0.00141	$2.3\cdot10^{-6}$
NMA(*trans*)/$\sigma_{\mathrm{ref,d}}$(T)	0.36668	-0.00112	$-6.8\cdot10^{-7}$	2.30550	-0.00674	$-5.5\cdot10^{-6}$	4.58169	-0.01437	$-7.5\cdot10^{-6}$
NMA(*trans*)/$\sigma_{\mathrm{ref,i}}$(T)	0.36668	-0.00112	$-6.8\cdot10^{-7}$	2.17017	-0.00603	$-6.3\cdot10^{-6}$	4.39573	-0.01340	$-8.6\cdot10^{-6}$
NMA(*trans*)/σ_{ref}(278.15K)	0.38800	-0.00124	$-5.7\cdot10^{-7}$	1.95971	-0.00596	$-3.9\cdot10^{-6}$	5.51605	-0.02078	$3.4\cdot10^{-6}$
NMA(*cis*)/$\sigma_{\mathrm{ref,d}}$(T)	0.32171	-0.00103	$-4.4\cdot10^{-7}$	2.50275	-0.00750	$-5.4\cdot10^{-6}$	5.14482	-0.01645	$-7.3\cdot10^{-6}$
NMA(*cis*)/$\sigma_{\mathrm{ref,i}}$(T)	0.32171	-0.00103	$-4.4\cdot10^{-7}$	2.36833	-0.00679	$-6.2\cdot10^{-6}$	4.95886	-0.01547	$-8.5\cdot10^{-6}$
NMA(*cis*)/σ_{ref}(278.15K)	0.34303	-0.00115	$-3.1\cdot10^{-7}$	2.15696	-0.00672	$-3.7\cdot10^{-6}$	6.07918	-0.02285	$3.6\cdot10^{-6}$
		^{1}H			^{13}C			^{15}N	
TMAO/δ_0 (calc)		3.155			66.064			108.975	
NMA(*trans*)/δ_0 (calc)		6.488			180.817			112.910	
NMA(*cis*)/δ_0 (calc)		5.953			183.728			109.941	

In this chapter, reference shielding constants for ^{1}H, ^{13}C and ^{15}N are presented, and different referencing methods are discussed. These shielding constants, together with the various pressure and temperature PMV corrections allow the use of the combined computational and NMR spectroscopic approach, presented for histamine in chapter 4.3, for the tautomer elucidation at various environmental conditions. This is essential for the clarification of the pressure/temperature dependence of the tautomeric equilibria of nucleic acid building blocks. In the following chapters, this is done with increasing complexity of the species of interest, starting with nucleobases up to nucleosides and nucleotides.

4.5 Tautomerism of nucleobases: Natural nucleobases, Hachimoji bases and derivatives

4.5.1 Investigated tautomers

In this chapter, the tautomerism of natural and non-natural nucleobases is investigated; the full set of Hachimoji bases is included. The considered nucleobases and tautomers are introduced in Figure 29 to Figure 33. To unambiguously identify each tautomer, in Table 20 a unique code is assigned to each tautomer, the same abbreviations are used for the tautomers of nucleobases, nucleosides and nucleotides. Natural nucleobases are divided into purine and pyrimidine bases; both structures and the corresponding atom numbering are shown in Figure 29 A and B. Especially the purine backbone can be modified in multiple ways; three modifications investigated in this work are shown in Figure 29 C-E.

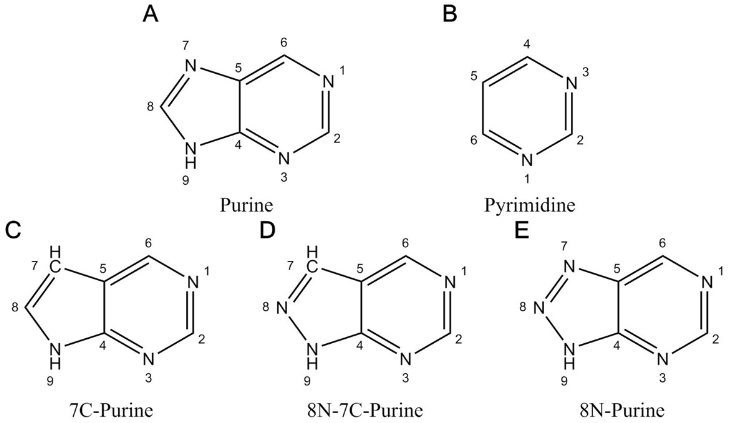


Figure 29: Backbone structures of the nucleobases investigated in this work. The backbones of the natural nucleobases; purine (A) and pyrimidine (B) are shown on top; the modified purines on the bottom, named by the changes in chemical structure with respect to the natural purine (C-E). In addition, the atom numbering is shown.

The natural purine bases, adenine and guanine, have three plausible tautomers each, neglecting the well-known tautomerism between the N9 and N7 positions due to the fact that the N9 is the binding site for the (desoxy-) ribose in the corresponding nucleosides. These tautomers are shown in Figure 30. Each of these tautomers exhibits a different hydrogen bonding pattern and is therefore able to form different base pairs. The relative stability of these tautomers is investigated alongside with the influence of extreme environmental conditions, high hydrostatic pressures and high temperatures, and the impact of changes in the purine backbone on these relative stabilities. Therefore, each of the tautomers is analysed with the natural purine scaffold as well as with the three modified versions from Figure 29 C-E.

In Figure 31, the analyzed tautomers of the natural pyrimidine nucleobases, uracil, thymine and cytosine, are depicted. Again, each of these tautomers has a unique pattern in terms of hydrogen bond donor and acceptor groups and is therefore able to undergo different base pairings. For the pyrimidine bases, the influence of extreme environmental conditions is explored, but no changes in the pyrimidine backbone are considered.

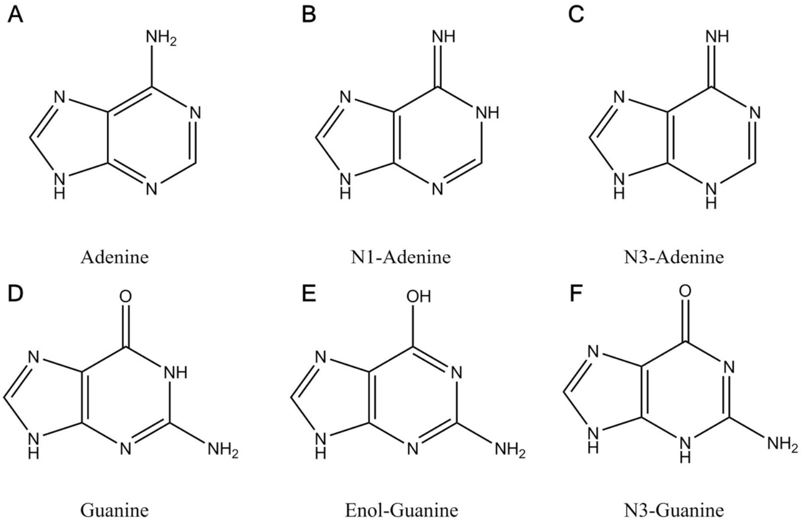


Figure 30: Tautomers of the natural purine nucleobases. Adenine (top row, A-C) and guanine (bottom row, D-E) tautomers investigated in this work. Structures in which the N7 is protonated instead of N9 are not considered (the N9 position is the (desoxy)-ribose binding site). The same tautomers are also used for the modified purine backbones shown in Figure 29 C-E.

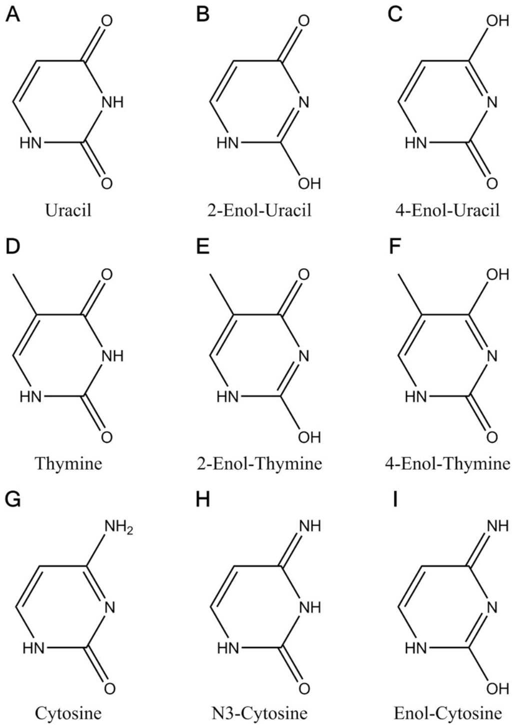


Figure 31: Tautomers of the natural pyrimidine nucleobases. Uracil (top row, A-C), thymine (middle row, D-F) and cytosine (bottom row, G-I) tautomers investigated in this work. Structures in which the N1 is deprotonated are not considered (the N1 position is the (desoxy)-ribose binding site).

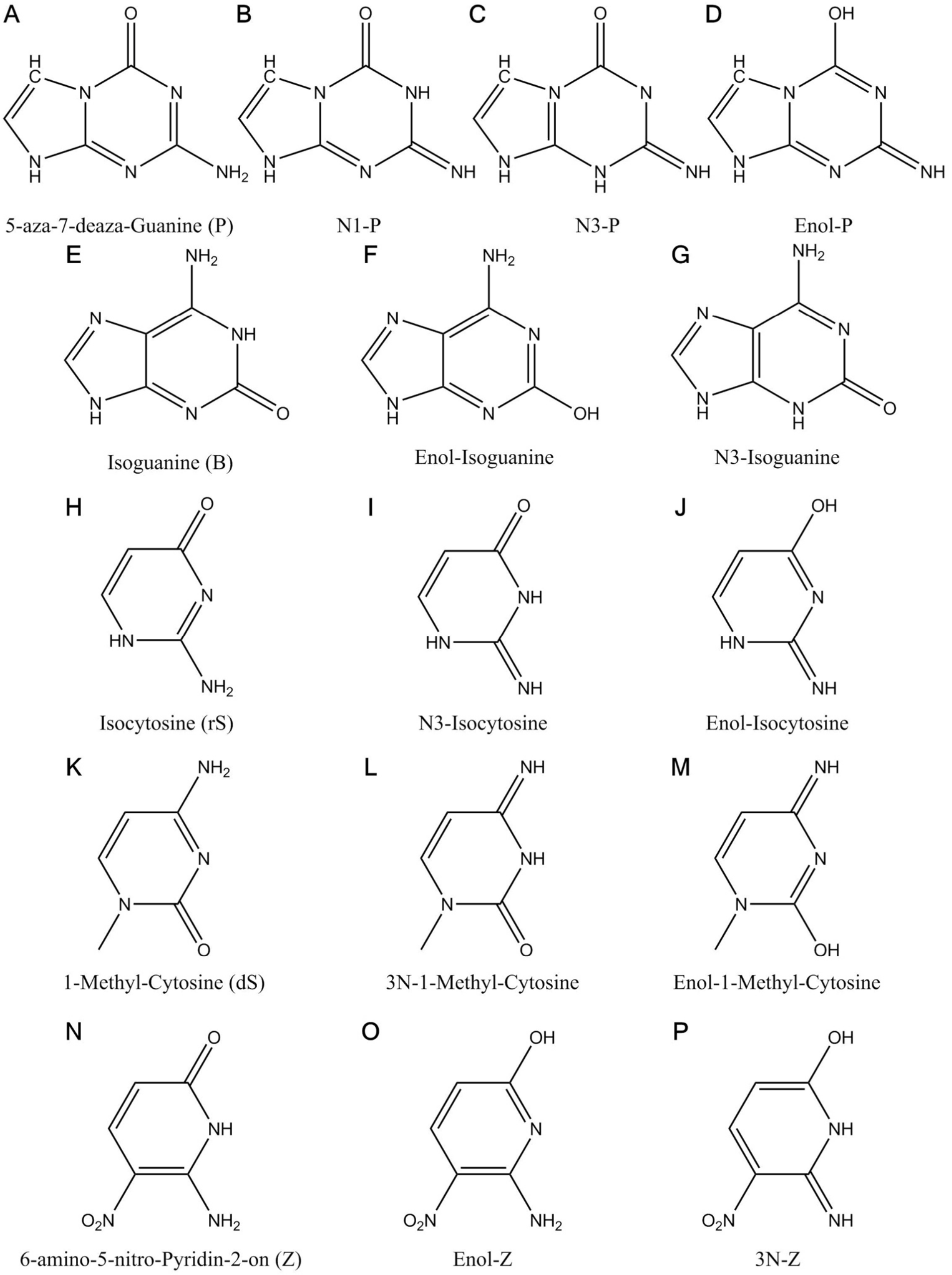
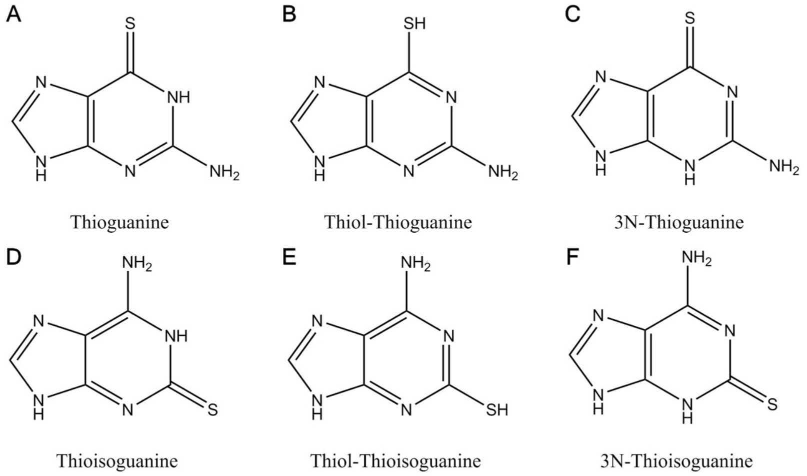


Figure 32: Tautomers of the Hachimoji nucleobases. 5-aza-7-deaza-Guanine (abbreviation P, A-D), isoguanine (abbreviation B, E-G), isocytosine (abbreviation rS, H-J), 1-methyl-cytosine (abbreviation dS, K-M) and 6-amino-5-nitro-pyridin-2-on (abbreviation Z, N-P) tautomers investigated in this work. Structures in which the N7 is protonated instead of N9 are not considered for P and B and structures with N1 deprotonation are not considered for rS (the N9 respectively N1 position is the (desoxy)-ribose binding site). The same tautomers of P and B are also used for the modified purine backbones shown in Figure 29 C-E.

The Hachimoji nucleobases are also investigated in this work, the three pyrimidines each have three possible tautomers with a unique hydrogen pattern, which is also true for the isoguanine, while the 5-aza-7-deazaguanine has four different tautomers; these tautomers are shown in Figure 32. For the isoguanine, also the different purine backbones are investigated. Besides, two more guanine derivatives are investigated: the guanine and the isoguanine, with the natural purine backbone but thiolated; their tautomers are depicted in Figure 33.



Thioguanine

Thiol-Thioguanine

3N-Thioguanine

Thioisoguanine

Thiol-Thioisoguanine

3N-Thioisoguanine

Figure 33: Tautomers of thiolated guanine nucleobases. Thioguanine (top row, A-C) and thioisoguanine (bottom row, D-E) tautomers investigated in this work. Structures in which the N7 is protonated instead of N9 are not considered (the N9 position is the (desoxy)-ribose binding site).

Table 20: Codes for all nucleic acid building block tautomers investigated in this work. Each code is assigned to a respective panel in Figure 30 to Figure 33. For tautomers with a purine backbone modification, the modifications shown in Figure 29 are assigned; if no modification is explicitly given, the standard nucleobase backbones (panels A and B) are used.

Code	Figure	Code	Figure	Code	Figure	Backbone	Code	Figure	Backbone
A	30 A	Enol-B	32 F	7C-A	30 A	29 C	8N-G	30 D	29 E
N1-A	30 B	N3-B	32 G	7C-N1-A	30 B	29 C	8N-Enol-G	30 E	29 E
N3-A	30 C	P	32 A	7C-N3-A	30 C	29 C	8N-N3-G	30 F	29 E
G	30 D	N1-P	32 B	8N-7C-A	30 A	29 D	7C-B	32 E	29 C
Enol-G	30 E	N3-P	32 C	8N-7C-N1-A	30 B	29 D	7C-Enol-B	32 F	29 C
N3-G	30 F	Enol-P	32 D	8N-7C-N3-A	30 C	29 D	7C-N3-B	32 G	29 C
C	31 G	Z	32 N	8N-A	30 A	29 E	8N-7C-B	32 E	29 D
Enol-C	31 I	N3-Z	32 O	8N-N1-A	30 B	29 E	8N-7C-Enol-B	32 F	29 D
N3-C	31 H	Enol-Z	32 P	8N-N3-A	30 C	29 E	8N-7C-N3-B	32 G	29 D
T	31 D	rS	32 H	7C-G	30 D	29 C	8N-B	32 E	29 E
2-Enol-T	31 E	Enol-rS	32 J	7C-Enol-G	30 E	29 C	8N-Enol-B	32 F	29 E
4-Enol-T	31 F	N3-rS	32 I	7C-N3-G	30 F	29 C	8N-N3-B	32 G	29 E
U	31 A	dS	32 K	8N-7C-G	30 D	29 D	Thio-B	33 D	-

Code	Figure	Code	Figure	Code	Figure	Backbone	Code	Figure	Backbone
2-Enol-U	31 B	Enol-dS	32 M	8N-7C-Enol-G	30 E	29 D	Thiol-Thio-B	33 E	-
4-Enol-U	31 C	N3-dS	32 L	8N-7C-N3-G	30 F	29 D	N3-Thio-B	33 F	-
B	32 E	Thio-G	33 A	Thiol-Thio-G	33 B	-	N3-Thio-G	33 C	-

4.5.2 Ambient conditions

For the investigation of nucleobase tautomerism at ambient conditions, not only EC-RISM is applied but also continuum solvation models like PCM. For both solvation models, the direct (Eqn. 199) and the indirect (Eqn. 200) path in the thermodynamic cycle are calculated, the latter using coupled-cluster calculations. Besides, the average of the EC-RISM approaches and the corresponding populations are calculated. The results for the natural nucleobases are given in Table 21.

Table 21: Results for the natural nucleobases at ambient conditions. The investigated tautomers are given in the first column; the reaction free energies ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/PCM (2nd), MP2/6-311+G(d,p)/PCM hydration free energy differences/CCSD(T)/cc-pVTZ gas phase reaction free energy (3rd), MP2/6-311+G(d,p)/EC-RISM (4th) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (5th) levels of theory are shown in columns 2-5. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. Literature values are taken from Ref. 300 (a), 301 (b), 302 (c), 303 (d) and 304 (e). The structures and results for the respective conformations are given in the SI part 05.

ΔG /	PCM	PCM	EC-RISM	EC-RISM	Average	Error	Average	Error	Literature
population		CCSD(T)		CCSD(T)	ΔG	ΔG	population	population	
A	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$1.99\cdot10^{-8}$	0.00
N1-A	11.23	11.62	8.83	9.01	8.92	0.04	$2.90\cdot10^{-7}$	$1.99\cdot10^{-8}$	6.3[b]/10.38[c]/12.91[a]
N3-A	20.83	18.15	16.71	14.03	15.37	0.60	$5.42\cdot10^{-12}$	$5.50\cdot10^{-12}$	4.4[b]/31.13[a]
G	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$1.25\cdot10^{-5}$	0.00
Enol-G	6.53	6.42	6.20	6.07	6.14	0.03	$3.18\cdot10^{-5}$	$1.47\cdot10^{-6}$	-6.47[c]/1.15[a]/4.95[d]/5.6[e]
N3-G	9.70	8.45	6.95	5.69	6.32	0.28	$2.33\cdot10^{-5}$	$1.10\cdot10^{-5}$	12.45[c]/3.32[d]/19.60[a]
C	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$9.15\cdot10^{-7}$	0.00
Enol-C	19.55	19.27	19.67	19.04	19.36	0.14	$6.00\cdot10^{-15}$	$2.00\cdot10^{-15}$	18.91[a]
N3-C	5.01	5.21	6.54	6.72	6.63	0.04	$1.38\cdot10^{-5}$	$9.15\cdot10^{-7}$	2.54[a]/3.69[c]
T	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$4.23\cdot10^{-8}$	0.00
2-Enol-T	14.44	13.54	12.19	11.13	11.66	0.24	$2.83\cdot10^{-9}$	$1.13\cdot10^{-9}$	12.45[a]
4-Enol-T	10.42	10.12	8.92	8.63	8.78	0.07	$3.69\cdot10^{-7}$	$4.12\cdot10^{-8}$	7.38[c]/18.45[a]
U	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$2.25\cdot10^{-7}$	0.00
2-Enol-U	14.85	13.97	12.55	11.58	12.07	0.22	$1.43\cdot10^{-9}$	$5.24\cdot10^{-10}$	19.14[a]
4-Enol-U	10.11	9.52	8.76	7.96	8.36	0.18	$7.48\cdot10^{-7}$	$2.22\cdot10^{-8}$	6.00[c]/11.76[a]

In general, the reaction free energies with respect to the Watson-Crick tautomer are larger using PCM compared to EC-RISM. Including high-level gas-phase energies reduces the energetic disadvantage of the minor tautomeric forms, but the Watson-Crick tautomers remain the most abundant with a fraction of nearly 1. Using EC-RISM, the N1 tautomer of adenine is stabilized by around 8.9 kcal/mol with respect to the Watson-Crick tautomer, the N3 tautomer by approximately 15.3 kcal/mol. This clearly shows the stability of the Watson-Crick adenine tautomer. The results from the literature differ in reaction free energies and energetic order of the minor tautomers, but in all cases, the Watson-Crick tautomer is the dominant species.

The stability of the Watson-Crick tautomer of guanine is comparable to the situation for adenine. The minor tautomers are penalized by more than 6 kcal/mol. Interestingly, the enol tautomer has an energetic difference of approximately 6.5 kcal/mol calculated using PCM and 6.1 kcal/mol with EC-RISM, nearly independent of the way used in the thermodynamic cycle. For the N3 tautomer, the results vary stronger between the different theoretical methods (5.7-9.7 kcal/mol reaction free energy). It is the second most abundant tautomer only at the highest level of theory (EC-RISM(CCSD(T))), leaving open questions concerning the ranking of the minor tautomers in free solution but clarifying the dominance of the Watson-Crick tautomer. Compared to results from literature, the data calculated in this work is more stable. In literature, the reaction free energies are widespread, the energetic order of the minor and even the main tautomer is unclear. Since TI calculations performed worse than EC-RISM (especially using CCSD(T) calculations) for the SAMPL2 dataset, they are not performed for the nucleobases presented in this chapter, but they are done (by coworkers, single- and double-topology calculations using MD and Monte-Carlo methods) in our recent publication (Ref. 9), yielding results consistent to EC-RISM and PCM with 6.7 respectively 7.3 kcal/mol reaction free energy for the enol- and 8.7 kcal/mol for the N3-tautomer. This, together with the experimental result of 5.6 kcal/mol[304] hints that the EC-RISM results here are more trustworthy than the strongly differing results from literature.

The cytosine Watson-Crick tautomer is clearly favored over the N3 tautomer (5.0-6.7 kcal/mol energetic difference) and the enol form, which is unpopulated with an energetic difference of 19.0-19.7 kcal/mol. This is consistent with literature.

Uracil and thymine behave in a similar way: The Watson-Crick tautomer is strongly favored by more than 7.9 kcal/mol, and the 2-enol is the least abundant one (reaction free energy of more than 11 kcal/mol). The stability of the Watson-Crick tautomer is, again, consistent with the literature, even though the energetic ordering of the minor thymine tautomers is ambiguous with respect to the literature. These findings confirm the Watson-Crick tautomers of all natural nucleobases to be the most stable ones with a large energetic difference, and therefore confirm the validity of the Watson-Crick pairs and their stability from a tautomeric point of view for ambient conditions.

The situation is more complicated for the Hachimoji nucleobases (Table 22). They are designed to expand the natural genetic code with four additional bases to increase the information density. They should therefore be comparably stable to the natural nucleobases regarding tautomers, to make the Hachimoji nucleic acids able to undergo Darwinian evolution in a comparable way while not heavily mutating because of switching hydrogen bonding patterns of the nucleobases upon tautomerization.

Table 22: Results for the Hachimoji nucleobases at ambient conditions. The investigated tautomers are given in the first column; the reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/PCM (2^nd^), MP2/6-311+G(d,p)/PCM hydration free energy differences/CCSD(T)/cc-pVTZ gas phase reaction free energy (3^rd^), MP2/6-311+G(d,p)/EC-RISM (4^th^) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (5^th^) levels of theory are shown in columns 2-5. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. Literature values are taken from Ref. 9, in parentheses the calculated results presented there are given, and Ref 305 (a). The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	PCM	PCM CCSD(T)	EC-RISM	EC-RISM CCSD(T)	Average ΔG	Error ΔG	Average population	Error population	Literature (Ref. 9)
B	0.00	0.00	0.00	0.00	0.00	0.00	0.746	0.034	0.00
Enol-B	0.96	1.90	1.57	2.53	2.05	0.21	0.023	0.009	1.4/6.7^a/6.8^a(6.6±0.2)
N3-B	0.05	-0.38	0.91	0.48	0.70	0.10	0.231	0.031	0.6/0.2^a(7.5±0.6)
P	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$3.92\cdot10^{-9}$	0.00
N1-P	10.56	10.22	10.76	10.24	10.50	0.12	$2.02\cdot10^{-8}$	$3.92\cdot10^{-9}$	(10.6)
N3-P	20.22	18.37	17.68	14.95	16.32	0.61	$1.10\cdot10^{-12}$	$1.13\cdot10^{-12}$	(16.9)
Enol-P	28.26	27.09	26.02	24.85	25.43	0.26	$<10^{-15}$	$<10^{-15}$	(25.5)
Z	0.00	0.00	0.00	0.00	0.00	0.00	0.998	$7.61\cdot10^{-4}$	0.00
N3-Z	20.33	20.42	19.01	19.04	19.03	0.01	$1.10\cdot10^{-14}$	$<10^{-15}$	(18.91)
Enol-Z	1.40	2.82	3.18	4.51	3.85	0.30	0.002	$7.61\cdot10^{-4}$	(3.8)
rS	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$4.59\cdot10^{-5}$	0.00
Enol-rS	12.34	11.96	14.20	13.29	13.75	0.21	$8.40\cdot10^{-11}$	$2.91\cdot10^{-11}$	(14.1)
N3-rS	3.43	2.70	5.57	4.80	5.19	0.17	$1.58\cdot10^{-4}$	$4.59\cdot10^{-5}$	(5.3)
dS	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$9.39\cdot10^{-6}$	-
Enol-dS	20.32	19.07	20.11	18.81	19.46	0.29	$5.00\cdot10^{-15}$	$3.00\cdot10^{-15}$	-
N3-dS	5.59	4.58	6.82	5.78	6.30	0.23	$2.40\cdot10^{-5}$	$9.39\cdot10^{-6}$	-

The isoguanine does not fulfil this requirement by having two minor tautomers within a small energetic difference to the main tautomer. The N3 tautomer has an energetic difference of -0.38-0.91 kcal/mol depending on the level of theory. Using PCM/CCSD(T), it is the main tautomer, using EC-RISM its fraction is 0.23, which should result in a lot of mispairings. For the least abundant tautomer, the energetic difference increases from 0.96-2.53 kcal/mol with regard to the Watson-Crick tautomer by going to higher levels of theory, but with a population of more than 2% this tautomer is by far more abundant than any minor tautomer of the natural nucleobases. This is also observed in literature,[306,307,308] resulting in investigations on purine backbone modifications of isoguanine to overcome the problem of multiple stable tautomers. This is also done in this work; the results are presented in Table 23.

In contrast, the Hachimoji P (5-aza-7-deaza-guanine), also a guanine derivative, has the highest reaction free energy to the minor tautomers with an energetic difference of approximately 10 kcal/mol and two other, energetically even higher tautomers, which are nearly not populated. It is therefore extremely tautomer stable, even compared to the natural nucleobases. In the Hachimoji P, the guanidine motif is, in contrast to isoguanine, preserved. This may contribute to the tautomer stability.

The Z (6-amino-5-nitro-pyridin-2-on) is a borderline case: while the N3 tautomer is clearly disfavored by approximately 20 kcal/mol, the enol form has a significant fraction of nearly 9% by using plain PCM. This fraction is shrinking by going to higher levels of theory, to 0.2% at the EC-RISM level. The associated energetic difference of 3.85 kcal/mol is slightly smaller than observed for the natural nucleobases, but this is not necessarily a bad sign for the use of this nucleobase in genetic codes. Overall, the main tautomer is very stable, and the slightly increased abundance of a minor tautomer may give the genetic code more evolutionary flexibility.

The rS (isocytosine), analogous to uracil only used in Hachimoji RNA, has a clearly disfavored enol tautomer and N3 tautomer which is significantly populated when using PCM (fraction of 0.003 with and 0.010 without high-level gas phase energies) but nearly unpopulated by using EC-RISM. Overall, the slightly smaller energetic differences between main and minor tautomers, compared to the cytosine, may be a reason why the cytosine prevailed in the evolution of nucleic acids.

The dS (1-methyl-cytosine) has a, consistent in all theoretical methods, unpopulated enol tautomer (approximately 20 kcal/mol reaction free energy) and a N3 minor tautomer with an energetic difference comparable to the natural nucleobase minor tautomers. The tautomer stability is not consistent for the Hachimoji nucleobases, while the P, dS and rS seem to be tautomer stable. Overall, an increased mutation rate is expected for an expanded eight letter genetic code mainly due to the tautomeric instability of the isoguanine (B) and, to a smaller extend, the 6-amino-5-nitro-pyridin-2-on (Z). Both will provide different hydrogen bonding patterns and not a single stable one due to their minor tautomers. In contrast to previous work,[9,306,307,308] a investigation of the tautomer stability of all Hachimoji nucleobases is presented here. In the next part, the influence of purine backbone modifications, which are also investigated in literature, to increase the tautomer stability of isoguanine, are studied.

The investigation of different purine backbones is not only important to investigate whether they can stabilize the main tautomer of isoguanine, it is also important to see the influence of these backbones onto the tautomerism of the natural nucleobases adenine and guanine. The natural nucleobases are used in DNA-encoded libraries where they are exposed to different reaction conditions and, especially the adenine, often depurinate at acidic conditions. [309,310,311] The use of modified purine backbones can help to overcome this issue. But as a prerequisite for the use of these nucleobases, the tautomer stability is important, since only tautomer stable nucleobases allow for the unambiguous reading of the barcode in an DNA-encoded library. The results of the investigation of the influence of variations in the purine backbone of natural, adenine and guanine, and non-natural, isoguanine, nucleobases are shown in Table 23.

Table 23: Results for the backbone variations of purine nucleobases at ambient conditions. The investigated tautomers are given in the first column, the reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/PCM (2nd), MP2/6-311+G(d,p)/PCM hydration free energy differences/CCSD(T)/cc-pVTZ gas phase reaction free energy (3rd), MP2/6-311+G(d,p)/EC-RISM (4th) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (5th) levels of theory are shown in columns 2-5. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. Literature values are taken from Ref. 9, in parentheses the calculated results presented there are given, and Ref. 312 (a). The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	PCM	PCM CCSD(T)	EC-RISM	EC-RISM CCSD(T)	Average ΔG	Error ΔG	Average population	Error population	Literature (Ref. 9)
7C-A	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$2.82\cdot10^{-7}$	-
7C-N1-A	10.89	9.70	9.13	7.87	8.50	0.28	$5.93\cdot10^{-7}$	$2.82\cdot10^{-7}$	-
7C-N3-A	19.61	15.80	16.74	12.63	14.69	0.92	$1.72\cdot10^{-11}$	$2.66\cdot10^{-11}$	-
8N-7C-A	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$3.48\cdot10^{-7}$	-
8N-7C-N1-A	9.25	9.16	7.31	7.13	7.22	0.04	$5.12\cdot10^{-6}$	$3.48\cdot10^{-7}$	-
8N-7C-N3-A	17.46	15.16	14.67	12.19	13.43	0.55	$1.43\cdot10^{-10}$	$1.34\cdot10^{-10}$	-

ΔG / population	PCM	PCM CCSD(T)	EC-RISM	EC-RISM CCSD(T)	Average ΔG	Error ΔG	Average population	Error population	Literature (Ref. 9)
8N-A	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$8.91\cdot10^{-7}$	-
8N-N1-A	9.04	8.97	7.24	6.88	7.06	0.08	$6.67\cdot10^{-6}$	$8.90\cdot10^{-7}$	-
8N-N3-A	18.04	15.19	14.93	12.13	13.53	0.63	$1.21\cdot10^{-10}$	$1.28\cdot10^{-10}$	-
7C-G	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$2.59\cdot10^{-5}$	-
7C-Enol-G	6.53	6.53	5.85	5.75	5.80	0.02	$5.58\cdot10^{-5}$	$2.07\cdot10^{-6}$	-
7C-N3-G	9.10	7.82	6.52	5.24	5.88	0.29	$4.92\cdot10^{-5}$	$2.38\cdot10^{-5}$	-
8N-7C-G	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$1.31\cdot10^{-5}$	-
8N-7C-Enol-G	8.14	7.83	7.19	6.80	6.99	0.09	$7.50\cdot10^{-6}$	$1.11\cdot10^{-6}$	-
8N-7C-N3-G	9.44	8.35	6.72	5.64	6.18	0.24	$2.94\cdot10^{-5}$	$1.20\cdot10^{-5}$	-
8N-G	0.00	0.00	0.00	0.00	0.00	0.00	>0.999	$7.85\cdot10^{-6}$	-
8N-Enol-G	8.67	7.94	7.60	6.86	7.23	0.17	$5.04\cdot10^{-6}$	$1.41\cdot10^{-6}$	-
8N-N3-G	9.69	8.56	7.15	6.01	6.58	0.25	$1.50\cdot10^{-5}$	$6.44\cdot10^{-6}$	-
7C-B	0.00	0.00	0.00	0.00	0.00	0.00	0.660	0.019	(0.0)
7C-Enol-B	1.25	3.45	1.88	4.07	2.97	0.49	0.004	0.004	(3.9)
7C-N3-B	-0.90	-1.10	0.50	0.30	0.40	0.04	0.336	0.018	(1.2)
8N-7C-B	0.00	0.00	0.00	0.00	0.00	0.00	0.558	0.030	(0.0)
8N-7C-Enol-B	1.44	3.20	2.32	4.08	3.20	0.39	0.003	0.002	(5.3)
8N-7C-N3-B	-1.17	-1.48	0.30	-0.01	0.14	0.07	0.439	0.030	(1.8)
8N-B	0.00	0.00	0.00	0.00	0.00	0.00	0.787	0.052	(0.0)
8N-Enol-B	0.74	3.08	1.39	3.72	2.55	0.52	0.011	0.010	3.04[a]/3.34[a](4.7)
8N-N3-B	-0.85	-0.11	0.44	1.17	0.80	0.16	0.203	0.047	0.04[a](2.0)

For adenine, the change of the N7 to a carbon only slightly decreases the preference of the major tautomer when considering the EC-RISM results, while with the exchange of the C8 with a nitrogen effect is somewhat more pronounced. Overall, these changes in the purine backbone influence the tautomer preferences of the natural adenine only slightly, the main tautomers remain the same and no significant fractions of minor forms occur. Some interesting phenomena can be observed for the guanine, the energetic ordering of the enol- and N3 tautomer switches. Changing the N7 to a carbon seems to favor the enol-tautomer, while the N8 increases the N3-preference. Nevertheless, the Watson-Crick tautomer is the dominant species. From the tautomer perspective, adenine and guanine can be used within DNA-encoded-libraries without hesitation.

The calculations on the problematic Hachimoji nucleobase isoguanine result in increased reaction free energies for the enol-tautomers regardless of the purine backbone modification used. This is consistent to the literature.[312] The N3-tautomers remain the problematic species. The EC-RISM calculations show no significant change in the stability of the Watson-Crick tautomer; in contrast, the N3-tautomer is slightly stabilized with the 7C-modification. The additional TI calculations presented in Ref. 9 indicate a small stabilization of the Watson-Crick tautomer, but still significant fractions of the N3-species. Therefore, we can conclude that there is still room for improvement of the Hachimoji B with respect to tautomer stability.

The last part of the investigation of nucleobases at ambient conditions is the examination of the thiolated guanines. Previous studies revealed that the thiolated guanine is comparably tautomer stable as the natural guanine.[313] The incorporation of the sulfur atom is another approach to increase the tautomer stability of isoguanine. The EC-RISM results indicate that this thiolation results in the N3-tautomer being the dominant form of thio-isoB with a minor fraction of the N1 form. But there is not only a switch in the dominant species upon thiolation, the thio-B is still not tautomer stable by having significant fractions of the minor tautomer. The use of the thio-B instead of isoguanine therefore would not increase the tautomer stability of Hachimoji nucleic acids but might enable alternative base pairs.

Table 24: Results for the thiolated guanines at ambient conditions. The investigated tautomers are given in the first column; the reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/PCM (2nd), MP2/6-311+G(d,p)/PCM hydration free energy differences/CCSD(T)/cc-pVTZ gas phase reaction free energy (3rd), MP2/6-311+G(d,p)/EC-RISM (4th) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (5th) levels of theory are shown in columns 2-5. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. Literature values are taken from Ref. 313. The structures and results for the respective conformations are given in the SI part 05.

Δ*G* /	PCM	PCM	EC-RISM	EC-RISM	Average	Error	Average	Error	Literature
population		CCSD(T)		CCSD(T)	Δ*G*	Δ*G*	population	population	
Thio-B	0.00	0.00	0.00	0.00	0.00	0.00	0.146	0.094	-
Thiol-Thio-B	-5.06	-2.73	-0.23	2.11	0.94	0.52	0.030	0.044	-
N3-Thio-B	-4.36	-2.38	-2.01	-0.03	-1.02	0.44	0.824	0.130	-
Thio-G	0.00	0.00	0.00	0.00	0.00	0.00	0.998	$3.04\cdot10^{-4}$	0.0
Thiol-Thio-G	5.72	3.39	6.18	3.62	4.90	0.57	$2.57\cdot10^{-4}$	$2.48\cdot10^{-4}$	10.5
N3-Thio-G	4.99	4.90	3.85	3.76	3.81	0.02	0.002	$5.63\cdot10^{-5}$	4.8/5.1

In this chapter, the tautomer stability of the natural nucleobases is shown, also the influence of a change in the purine backbone of adenine and guanine and the thiolation of guanine is investigated. All these modifications affect the tautomer stability of the natural species only slightly. In contrast, the Hachimoji nucleobases are not completely tautomer stable, the most problematic species is the isoguanine, with a major fraction of the N3-tautomer, which is not sufficiently stabilized via purine backbone modifications or thiolation, which may lead to increased mutation events in the Hachimoji code. In the next chapters, the influence of high pressures and temperature variations on the tautomeric equilibria is investigated. This is of special interest since high temperatures and pressures are mimicking early life conditions, and the influence of these environmental conditions on the tautomeric equilibria, and therefore mutation rates in the genetic code, may have played an important role in the evolution.

4.5.3 High hydrostatic pressure conditions

The tautomerism of nucleic acid building blocks is worth an investigation not only at ambient conditions since early life conditions are important from an evolutionary point of view. At deep oceanic black smokers, high hydrostatic pressures occur. The rule of thumb is that there is 1 bar additional hydrostatic pressure in oceans of the earth for each 10 meters of depth, resulting in 1 kbar pressure at the Mariana trench. Hypothetically, hydrostatic pressures can go up to 10 kbar at 300 K until water freezes, therefore these two pressures have been chosen to be presented here, but the calculations are performed at all pressures given in the computational details, and the full results are presented in the SI part 05. The pressure dependent trends are monotonous (with small exceptions at minor, unpopulated tautomers), so the influence of hydrostatic pressure on the tautomerism of nucleobases is covered by discussing two pressures like it is done here.

While EC-RISM is capable to incorporate high hydrostatic pressure effects, continuum electrostatic models are usually not able to do this, so no PCM results are discussed here. The reaction free energies of the natural nucleobases (Table 25) decrease upon pressurization for most minor tautomers (the exceptions are the enol-guanine, enol-cytosine and N3-cytosine), resulting in increased populations of the minor tautomers, even though the smallest energetic difference of a minor tautomer is approximately 5.8 kcal/mol. The largest fraction of a minor tautomer is the N3-tautomer of guanine with a sub ‰ fraction. Overall, the number of alternative hydrogen bonding patterns is increased slightly upon pressurization, which may lead to more mutations and therefore enhanced evolutionary dynamics at high pressures.

Table 25: Results for the natural nucleobases at 1 kbar and 10 kbar (at 298.15 K). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (1 kbar) and 9th (10 kbar)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (1 kbar)/ 10th (10 kbar)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (1 kbar) and 11-14 (10 kbar). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 1 bar	EC-RISM 1 kbar	EC-RISM/CCSD(T) 1 kbar	Average ΔG 1 kbar	Error ΔG 1 kbar	Average pop. 1 kbar	Error pop. 1 kbar	EC-RISM 10 kbar	EC-RISM/CCSD(T) 10 kbar	Average ΔG 10 kbar	Error ΔG 10 kbar	Average pop. 10 kbar	Error pop. 10 kbar
A	>0.999	0.00	0.00	0.00	0.00	>0.999	$2.16\cdot10^{-8}$	0.00	0.00	0.00	0.00	>0.999	$2.62\cdot10^{-8}$
N1-A	$2.90\cdot10^{-7}$	8.77	8.94	8.86	0.04	$3.23\cdot10^{-7}$	$2.16\cdot10^{-8}$	8.59	8.75	8.67	0.04	$4.42\cdot10^{-7}$	$2.61\cdot10^{-8}$
N3-A	$5.42\cdot10^{-12}$	16.58	13.89	15.24	0.60	$6.79\cdot10^{-12}$	$6.89\cdot10^{-12}$	16.13	13.44	14.79	0.60	$1.45\cdot10^{-11}$	$1.47\cdot10^{-11}$
G	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.44\cdot10^{-5}$	0.00	0.00	0.00	0.00	>0.999	$2.61\cdot10^{-5}$
Enol-G	$3.18\cdot10^{-5}$	6.27	6.14	6.20	0.03	$2.84\cdot10^{-5}$	$1.31\cdot10^{-6}$	6.59	6.46	6.53	0.03	$1.64\cdot10^{-5}$	$7.97\cdot10^{-7}$
N3-G	$2.33\cdot10^{-5}$	6.85	5.59	6.22	0.28	$2.76\cdot10^{-5}$	$1.31\cdot10^{-5}$	6.45	5.20	5.83	0.28	$5.37\cdot10^{-5}$	$2.54\cdot10^{-5}$
C	>0.999	0.00	0.00	0.00	0.00	>0.999	$8.66\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$6.41\cdot10^{-7}$
Enol-C	$6.00\cdot10^{-15}$	19.69	19.04	19.36	0.14	$6.00\cdot10^{-15}$	$2.00\cdot10^{-15}$	19.88	19.16	19.52	0.16	$5.00\cdot10^{-15}$	$1.00\cdot10^{-15}$
N3-C	$1.38\cdot10^{-5}$	6.57	6.75	6.66	0.04	$1.31\cdot10^{-5}$	$8.66\cdot10^{-7}$	6.74	6.91	6.83	0.04	$9.89\cdot10^{-6}$	$6.41\cdot10^{-7}$
T	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.43\cdot10^{-8}$	0.00	0.00	0.00	0.00	>0.999	$4.86\cdot10^{-8}$
2-Enol-T	$2.83\cdot10^{-9}$	12.15	11.08	11.62	0.24	$3.06\cdot10^{-9}$	$1.24\cdot10^{-9}$	12.07	10.94	11.51	0.25	$3.69\cdot10^{-9}$	$1.58\cdot10^{-9}$
4-Enol-T	$3.69\cdot10^{-7}$	8.90	8.60	8.75	0.07	$3.84\cdot10^{-7}$	$4.30\cdot10^{-8}$	8.87	8.56	8.71	0.07	$4.10\cdot10^{-7}$	$4.70\cdot10^{-8}$
U	>0.999	0.00	0.00	0.00	0.00	>0.999	$2.41\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$2.88\cdot10^{-7}$
2-Enol-U	$1.43\cdot10^{-9}$	12.51	11.53	12.02	0.22	$1.54\cdot10^{-9}$	$5.68\cdot10^{-9}$	12.45	11.42	11.93	0.23	$1.79\cdot10^{-9}$	$6.92\cdot10^{-10}$
4-Enol-U	$7.48\cdot10^{-7}$	8.74	7.92	8.33	0.18	$7.82\cdot10^{-7}$	$2.40\cdot10^{-7}$	8.73	7.83	8.28	0.20	$8.52\cdot10^{-7}$	$2.88\cdot10^{-7}$

The Hachimoji B minor tautomers are destabilized at higher pressures (Table 26), but at 10 kbar, the enol- and N3-tautomer are still occurring in large fractions, leaving this base problematic by exhibiting three different hydrogen bonding patterns which may lead to a lot of mispairings. In contrast, the Hachimoji P minor tautomers become stabilized at high pressures; the N3-tautomer is more stabilized than the other minor tautomers. Overall, these changes upon pressurization are not affecting the main tautomer, it is still the only populated one with a fraction of nearly 1. The Hachimoji P is very stable under high pressures. The Z is nearly unaffected by pressure. At 10 kbar, the enol tautomer is still populated significantly stronger than every natural nucleobase minor tautomer. The dS and rS main tautomers are stabilized at higher pressures: the dS shows no significant occurrence of minor tautomers and is strongly preferred, the rS is stabilized in a way that it has nearly the same tautomer stability as guanine at 10 kbar.

Table 26: Results for the Hachimoji nucleobases at 1 kbar and 10 kbar (at 298.15 K). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (1 kbar) and 9th (10 kbar)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (1 kbar)/ 10th (10 kbar)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (1 kbar) and 11-14 (10 kbar). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 1 bar	EC-RISM 1 kbar	EC-RISM/CCSD(T) 1 kbar	Average ΔG 1 kbar	Error ΔG 1 kbar	Average pop. 1 kbar	Error pop. 1 kbar	EC-RISM 10 kbar	EC-RISM/CCSD(T) 10 kbar	Average ΔG 10 kbar	Error ΔG 10 kbar	Average pop. 10 kbar	Error pop. 10 kbar
B	0.746	0.00	0.00	0.00	0.00	0.754	0.033	0.00	0.00	0.00	0.00	0.784	0.029
Enol-B	0.023	1.69	2.65	2.17	0.21	0.019	0.008	2.23	3.19	2.71	0.22	0.008	0.003
N3-B	0.231	0.93	0.50	0.71	0.10	0.226	0.030	1.00	0.57	0.79	0.10	0.208	0.027
P	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.09\cdot10^{-9}$	0.00	0.00	0.00	0.00	>0.999	$4.37\cdot10^{-9}$
N1-P	$2.02\cdot10^{-8}$	10.74	10.22	10.48	0.12	$2.08\cdot10^{-8}$	$4.09\cdot10^{-9}$	10.74	10.20	10.47	0.12	$2.12\cdot10^{-8}$	$4.37\cdot10^{-9}$
N3-P	$1.10\cdot10^{-12}$	17.55	14.79	16.17	0.62	$1.41\cdot10^{-12}$	$1.46\cdot10^{-12}$	17.10	14.23	15.67	0.64	$3.27\cdot10^{-12}$	$3.54\cdot10^{-12}$
Enol-P	$<10^{-15}$	25.98	24.81	25.39	0.26	$<10^{-15}$	$<10^{-15}$	25.98	24.81	25.39	0.26	$<10^{-15}$	$<10^{-15}$
Z	0.998	0.00	0.00	0.00	0.00	0.999	$6.38\cdot10^{-4}$	0.00	0.00	0.00	0.00	>0.999	$2.84\cdot10^{-4}$
N3-Z	$1.10\cdot10^{-14}$	19.00	19.03	19.01	0.01	$1.20\cdot10^{-14}$	$<10^{-15}$	19.06	19.09	19.08	0.01	$1.00\cdot10^{-14}$	$<10^{-15}$
Enol-Z	0.002	3.28	4.61	3.95	0.30	0.001	$6.38\cdot10^{-4}$	3.77	5.06	4.41	0.29	$5.81\cdot10^{-4}$	$2.84\cdot10^{-4}$
rS	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.21\cdot10^{-5}$	0.00	0.00	0.00	0.00	>0.999	$2.85\cdot10^{-5}$
Enol-rS	$8.40\cdot10^{-11}$	14.27	13.32	13.80	0.21	$7.69\cdot10^{-11}$	$2.76\cdot10^{-11}$	14.64	13.56	14.10	0.24	$4.63\cdot10^{-11}$	$1.89\cdot10^{-11}$
N3-rS	$1.58\cdot10^{-4}$	5.62	4.85	5.24	0.17	$1.45\cdot10^{-4}$	$4.21\cdot10^{-5}$	5.86	5.09	5.47	0.17	$9.74\cdot10^{-5}$	$2.85\cdot10^{-5}$
dS	>0.999	0.00	0.00	0.00	0.00	>0.999	$9.03\cdot10^{-6}$	0.00	0.00	0.00	0.00	>0.999	$7.03\cdot10^{-6}$
Enol-dS	$5.00\cdot10^{-15}$	20.11	18.81	19.46	0.29	$5.00\cdot10^{-15}$	$3.00\cdot10^{-15}$	20.24	18.94	19.59	0.29	$4.00\cdot10^{-15}$	$2.00\cdot10^{-15}$
N3-dS	$2.40\cdot10^{-5}$	6.84	5.81	6.33	0.23	$2.31\cdot10^{-5}$	$9.03\cdot10^{-6}$	7.00	5.96	6.48	0.23	$1.79\cdot10^{-5}$	$7.03\cdot10^{-6}$

The natural nucleobases are also marginally affected by pressure when the alternative purine backbones are used (Table 27). The main tautomer of adenine is destabilized under high pressures regardless of the used backbone modifications. Overall, the 8N-modified adenines are more unstable than the natural one, but the reaction free energies are still high with approx. 7 kcal/mol at 10 kbar. In case of guanine, the enol gets destabilized by pressure, and the N3-tautomer is stabilized with all modifications. The 7C-guanine shows the largest minor tautomer fraction with 0.1 ‰. The opposite trend can be observed for the problematic N3-

tautomer of the isoguanine. The N1 tautomer gets stabilized by pressure regardless of the backbone modification, but even at 10 kbar, there is a large fraction of the N3 tautomer, which makes these nucleobases not well suited for a tautomer stable genetic code at the pressure range investigated here.

Table 27: Results for the backbone variations of the purine nucleobases at 1 kbar and 10 kbar (at 298.15 K). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (1 kbar) and 9th (10 kbar)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (1 kbar)/ 10th (10 kbar)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (1 kbar) and 11-14 (10 kbar). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 1 bar	EC-RISM 1 kbar	EC-RISM/CCSD(T) 1 kbar	Average ΔG 1 kbar	Error ΔG 1 kbar	Average pop. 1 kbar	Error pop. 1 kbar	EC-RISM 10 kbar	EC-RISM/CCSD(T) 10 kbar	Average ΔG 10 kbar	Error ΔG 10 kbar	Average pop. 10 kbar	Error pop. 10 kbar
7C-A	>0.999	0,00	0,00	0,00	0,00	>0.999	$3.10\cdot10^{-7}$	0,00	0,00	0,00	0,00	>0.999	$4.14\cdot10^{-7}$
7C-N1-A	$5.93\cdot10^{-7}$	9,07	7,81	8,44	0,28	$6.51\cdot10^{-7}$	$3.10\cdot10^{-7}$	8,91	7,64	8,27	0,28	$8.61\cdot10^{-7}$	$4.14\cdot10^{-7}$
7C-N3-A	$1.72\cdot10^{-11}$	16,62	12,51	14,56	0,92	$2.11\cdot10^{-11}$	$3.28\cdot10^{-11}$	16,20	12,09	14,15	0,92	$4.26\cdot10^{-11}$	$6.61\cdot10^{-11}$
8N-7C-A	>0.999	0,00	0,00	0,00	0,00	>0.999	$3.83\cdot10^{-7}$	0,00	0,00	0,00	0,00	>0.999	$5.12\cdot10^{-7}$
8N-7C-N1-A	$5.12\cdot10^{-6}$	7,26	7,08	7,17	0,04	$5.56\cdot10^{-6}$	$3.83\cdot10^{-7}$	7,13	6,93	7,03	0,04	$7.03\cdot10^{-6}$	$5.11\cdot10^{-7}$
8N-7C-N3-A	$1.43\cdot10^{-10}$	14,56	12,08	13,32	0,56	$1.73\cdot10^{-10}$	$1.62\cdot10^{-10}$	14,20	11,70	12,95	0,56	$3.21\cdot10^{-10}$	$3.02\cdot10^{-10}$
8N-A	>0.999	0,00	0,00	0,00	0,00	>0.999	$1.02\cdot10^{-6}$	0,00	0,00	0,00	0,00	>0.999	$1.61\cdot10^{-6}$
8N-N1-A	$6.67\cdot10^{-6}$	7,18	6,81	6,99	0,08	$7.47\cdot10^{-6}$	$1.02\cdot10^{-6}$	6,99	6,59	6,79	0,09	$1.06\cdot10^{-5}$	$1.61\cdot10^{-6}$
8N-N3-A	$1.21\cdot10^{-10}$	14,81	12,00	13,40	0,63	$1.50\cdot10^{-10}$	$1.58\cdot10^{-10}$	14,38	11,58	12,98	0,63	$3.06\cdot10^{-10}$	$3.24\cdot10^{-10}$
7C-G	>0.999	0,00	0,00	0,00	0,00	>0.999	$2.98\cdot10^{-5}$	0,00	0,00	0,00	0,00	>0.999	$5.15\cdot10^{-5}$
7C-Enol-G	$5.58\cdot10^{-5}$	5,92	5,82	5,87	0,02	$4.98\cdot10^{-5}$	$1.97\cdot10^{-6}$	6,25	6,11	6,18	0,03	$2.96\cdot10^{-5}$	$1.56\cdot10^{-6}$
7C-N3-G	$4.92\cdot10^{-5}$	6,43	5,14	5,78	0,29	$5.75\cdot10^{-5}$	$2.78\cdot10^{-5}$	6,08	4,80	5,44	0,29	$1.03\cdot10^{-4}$	$4.99\cdot10^{-5}$
8N-7C-G	>0.999	0,00	0,00	0,00	0,00	>0.999	$1.53\cdot10^{-5}$	0,00	0,00	0,00	0,00	>0.999	$2.73\cdot10^{-5}$
8N-7C-Enol-G	$7.50\cdot10^{-6}$	7,26	6,86	7,06	0,09	$6.69\cdot10^{-6}$	$1.00\cdot10^{-6}$	7,59	7,16	7,37	0,10	$3.93\cdot10^{-6}$	$6.46\cdot10^{-7}$
8N-7C-N3-G	$2.94\cdot10^{-5}$	6,63	5,54	6,08	0,24	$3.47\cdot10^{-5}$	$1.42\cdot10^{-5}$	6,25	5,17	5,71	0,24	$6.51\cdot10^{-5}$	$2.66\cdot10^{-5}$
8N-G	>0.999	0,00	0,00	0,00	0,00	>0.999	$8.98\cdot10^{-6}$	0,00	0,00	0,00	0,00	>0.999	$1.63\cdot10^{-5}$
8N-Enol-G	$5.04\cdot10^{-6}$	7,67	6,93	7,30	0,17	$4.49\cdot10^{-6}$	$1.26\cdot10^{-6}$	8,00	7,25	7,63	0,17	$2.56\cdot10^{-6}$	$7.23\cdot10^{-7}$
8N-N3-G	$1.50\cdot10^{-5}$	7,04	5,90	6,47	0,25	$1.80\cdot10^{-5}$	$7.73\cdot10^{-6}$	6,63	5,49	6,06	0,25	$3.63\cdot10^{-5}$	$1.56\cdot10^{-5}$
7C-B	0.660	0,00	0,00	0,00	0,00	0,670	0,018	0,00	0,00	0,00	0,00	0,705	0,016
7C-Enol-B	0.004	2,00	4,20	3,10	0,49	0,004	0,003	2,55	4,74	3,64	0,49	0,002	0,001
7C-N3-B	0.336	0,53	0,33	0,43	0,04	0,327	0,018	0,62	0,42	0,52	0,04	0,293	0,016
8N-7C-B	0.558	0,00	0,00	0,00	0,00	0,574	0,030	0,00	0,00	0,00	0,00	0,633	0,028
8N-7C-Enol-B	0.003	2,46	4,22	3,34	0,39	0,002	0,001	3,05	4,80	3,93	0,39	$8.37\cdot10^{-4}$	$5.90\cdot10^{-4}$
8N-7C-N3-B	0.439	0,34	0,02	0,18	0,07	0,424	0,030	0,48	0,17	0,32	0,07	0,366	0,028
8N-B	0.787	0,00	0,00	0,00	0,00	0,799	0,049	0,00	0,00	0,00	0,00	0,845	0,038
8N-Enol-B	0.011	1,52	3,86	2,69	0,52	0,009	0,008	2,08	4,42	3,25	0,52	0,004	0,003
8N-N3-B	0.203	0,48	1,21	0,84	0,16	0,192	0,045	0,65	1,39	1,02	0,16	0,151	0,036

Like the pressure dependent trends for the guanine and isoguanine with the purine backbone modifications, the trends for the thiolated guanine and isoguanine (Table 28) are in line with the ones observed for the original species. The N3-tautomer is still the main species of the thioisoguanine at 10 kbar, but the Watson-Crick tautomer becomes more stabilized upon pressurization.

Table 28: Results for the thiolated guanines at 1 kbar and 10 kbar (at 298.15 K). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (1 kbar) and 9th (10 kbar)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (1 kbar)/ 10th (10 kbar)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (1 kbar) and 11-14 (10 kbar). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 1 bar	EC-RISM 1 kbar	EC-RISM/CCSD(T) 1 kbar	Average ΔG 1 kbar	Error ΔG 1 kbar	Average pop. 1 kbar	Error pop. 1 kbar	EC-RISM 10 kbar	EC-RISM/CCSD(T) 10 kbar	Average ΔG 10 kbar	Error ΔG 10 kbar	Average pop. 10 kbar	Error pop. 10 kbar
Thio-B	0.146	0.00	0.00	0.00	0.00	0.167	0.104	0.00	0.00	0.00	0.00	0.268	0.147
Thiol-Thio-B	0.030	-0.01	2.33	1.16	0.52	0.024	0.035	0.84	3.18	2.01	0.52	0.009	0.013
N3-Thio-B	0.824	-1.92	0.05	-0.93	0.44	0.809	0.132	-1.58	0.40	-0.59	0.44	0.723	0.155
Thio-G	0.998	0.00	0.00	0.00	0.00	0.998	$2.74 \cdot 10^{-4}$	0.00	0.00	0.00	0.00	0.998	$2.01 \cdot 10^{-4}$
Thiol-Thio-G	$2.57 \cdot 10^{-4}$	6.26	3.70	4.98	0.57	$2.22 \cdot 10^{-4}$	$2.15 \cdot 10^{-4}$	6.59	4.02	5.30	0.58	$1.29 \cdot 10^{-4}$	$1.26 \cdot 10^{-4}$
N3-Thio-G	0.002	3.82	3.73	3.77	0.02	0.002	$5.96 \cdot 10^{-5}$	3.67	3.58	3.62	0.02	0.002	$7.63 \cdot 10^{-5}$

The natural nucleobases are evolutionary very well chosen, since they are not only tautomer stable at ambient conditions but also under the high-pressure conditions investigated here, even with small modifications like incorporation of sulfur instead of oxygen or by changing the purine backbone of adenine and guanine. The problematic minor tautomers of the Hachimoji B and Z bases are sensitive to higher pressures, they become destabilized. But the isoguanine has large fractions of minor tautomers even with backbone variations. The differences between the natural nucleobases, especially the guanine and the Hachimoji bases, especially the isoguanine, are remarkable. The natural nucleobases tend to destabilize at high pressures while the Hachimoji bases are unaffected or overall stabilized by pressure. The structurally very similar guanine and isoguanine show the strongest of these trends. The highest pressure investigated here is 10 kbar, at this pressure the guanine is still relatively stable and the isoguanine unstable, but at much higher pressures, which may occur on exoplanets, it is possible that the Hachimoji bases are tautomer stable. This hints that non-canonical nucleobases can be a valid alternative to the natural genetic code.

4.5.4 High temperature conditions

While natural nucleobases remain tautomer-stable at high hydrostatic pressure conditions, the influence of temperature is also needed to investigate since the genetic code is universal for life on earth which can exist at strongly varying temperatures. Since the calculations are done in liquid water at ambient pressures, this investigation has to be done in the temperature range of liquid water. The lowest temperature chosen here is 278.15 K and the highest one 363.15 K, the calculations are done at all temperatures given in the computational details and can be found in detail in the SI part 05. The extreme temperatures chosen here should give the most insight into the tautomerism of nucleobases under different temperatures, since the deviation from the ambient conditions is as large as possible. It is important when calculating the tautomer fractions at various temperatures not only to consider the reaction free energies since also the Boltzmann factor is temperature dependent.

Table 29: Results for the natural nucleobases at 278.15 K and 363.15 K (at 1bar). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (278.15 K) and 9th (363.15 K)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (278.15 K)/ 10th (363.15 K)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (278.15 K) and 11-14 (363.15 K). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
A	>0.999	0.00	0.00	0.00	0.00	>0.999	$8.20 \cdot 10^{-9}$	0.00	0.00	0.00	0.00	>0.999	$1.83 \cdot 10^{-7}$
N1-A	$2.90 \cdot 10^{-7}$	8.76	8.95	8.86	0.04	$1.10 \cdot 10^{-7}$	$8.20 \cdot 10^{-9}$	8.91	9.07	8.99	0.03	$3.89 \cdot 10^{-6}$	$1.83 \cdot 10^{-7}$
N3-A	$5.42 \cdot 10^{-12}$	16.60	13.91	15.26	0.60	$1.03 \cdot 10^{-12}$	$1.12 \cdot 10^{-12}$	16.99	14.31	15.65	0.60	$3.81 \cdot 10^{-10}$	$3.17 \cdot 10^{-10}$
G	>0.999	0.00	0.00	0.00	0.00	>0.999	$6.87 \cdot 10^{-6}$	0.00	0.00	0.00	0.00	>0.999	$4.88 \cdot 10^{-5}$

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
Enol-G	$3.18\cdot10^{-5}$	6.18	6.06	6.12	0.03	$1.56\cdot10^{-5}$	$7.39\cdot10^{-7}$	6.01	5.88	5.94	0.03	$2.65\cdot10^{-4}$	$1.10\cdot10^{-5}$
N3-G	$2.33\cdot10^{-5}$	6.88	5.63	6.26	0.28	$1.21\cdot10^{-5}$	$6.13\cdot10^{-6}$	7.29	6.04	6.66	0.28	$9.75\cdot10^{-5}$	$3.78\cdot10^{-5}$
C	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.02\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$7.03\cdot10^{-6}$
Enol-C	$6.00\cdot10^{-15}$	19.68	19.06	19.37	0.14	$1.00\cdot10^{-15}$	$<10^{-15}$	19.64	19.00	19.32	0.14	$2.36\cdot10^{-12}$	$4.67\cdot10^{-13}$
N3-C	$1.38\cdot10^{-5}$	6.60	6.78	6.69	0.04	$5.58\cdot10^{-6}$	$4.02\cdot10^{-7}$	6.35	6.52	6.43	0.04	$1.34\cdot10^{-4}$	$7.03\cdot10^{-6}$
T	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.66\cdot10^{-8}$	0.00	0.00	0.00	0.00	>0.999	$4.42\cdot10^{-7}$
2-Enol-T	$2.83\cdot10^{-9}$	12.12	11.07	11.59	0.23	$7.78\cdot10^{-10}$	$3.29\cdot10^{-10}$	12.45	11.36	11.90	0.24	$6.85\cdot10^{-8}$	$2.31\cdot10^{-8}$
4-Enol-T	$3.69\cdot10^{-7}$	8.87	8.58	8.73	0.06	$1.39\cdot10^{-7}$	$1.63\cdot10^{-8}$	9.10	8.77	8.94	0.07	$4.19\cdot10^{-6}$	$4.19\cdot10^{-7}$
U	>0.999	0.00	0.00	0.00	0.00	>0.999	$9.24\cdot10^{-8}$	0.00	0.00	0.00	0.00	>0.999	$1.99\cdot10^{-6}$
2-Enol-U	$1.43\cdot10^{-9}$	12.48	11.52	12.00	0.21	$3.72\cdot10^{-10}$	$1.44\cdot10^{-10}$	12.80	11.80	12.30	0.22	$3.96\cdot10^{-8}$	$1.22\cdot10^{-8}$
4-Enol-U	$7.48\cdot10^{-7}$	8.71	7.93	8.32	0.18	$2.90\cdot10^{-7}$	$9.22\cdot10^{-8}$	8.88	8.07	8.48	0.18	$7.90\cdot10^{-6}$	$1.99\cdot10^{-6}$

The minor tautomers of adenine show increasing reaction free energies with temperature (Table 29), but due to the Boltzmann factor, their fractions are also slightly increasing. The same behavior is observed for the N3-guanine; the enol-guanine reaction free energy shows a non-linear behavior, it is decreased, compared to the 298.15 K value, at high and low temperatures, resulting in a 0.27 ‰ fraction at 363.15 K. This non-linear behavior is also recognized for the cytosine minor tautomers, with the largest fraction being 1.34 ‰ at 363.15 K. Both of the other pyrimidine bases, thymine and uracil, react like adenine at increasing temperatures: Although the reaction free energies of the minor tautomers increase, their fractions increase simultaneously due to the Boltzmann factor. The dominance of the natural nucleobase Watson-Crick tautomers is not affected from an energetic point of view, but the minor tautomer fractions are slightly increasing. The Watson-Crick tautomers are the most abundant species in the temperature range investigated here, so they are stable over a broad temperature range (especially considering the errors in populations of the minor tautomers). Still, the slightly increased minor tautomer fractions at high temperatures, may lead to increased mutation rates, which would result in a faster evolution at early life conditions.

Table 30: Results for the Hachimoji nucleobases at 278.15 K and 363.15 K (at 1bar). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (278.15 K) and 9th (363.15 K)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (278.15 K)/ 10th (363.15 K)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (278.15 K) and 11-14 (363.15 K). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
B	0.746	0.00	0.00	0.00	0.00	0.770	0.033	0.00	0.00	0.00	0.00	0.659	0.037
Enol-B	0.023	1.67	2.62	2.15	0.21	0.016	0.007	1.16	2.13	1.64	0.22	0.068	0.021
N3-B	0.231	0.92	0.49	0.71	0.10	0.214	0.031	0.85	0.42	0.63	0.10	0.273	0.032
P	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.15\cdot10^{-9}$	0.00	0.00	0.00	0.00	>0.999	$7.98\cdot10^{-8}$
N1-P	$2.02\cdot10^{-8}$	10.75	10.25	10.50	0.11	$5.64\cdot10^{-9}$	$1.15\cdot10^{-9}$	10.77	10.23	10.50	0.12	$4.78\cdot10^{-7}$	$7.97\cdot10^{-8}$
N3-P	$1.10\cdot10^{-12}$	17.60	14.86	16.23	0.61	$1.78\cdot10^{-13}$	$1.98\cdot10^{-13}$	18.04	15.37	16.71	0.60	$8.81\cdot10^{-11}$	$7.29\cdot10^{-11}$
Enol-P	$<10^{-15}$	25.95	24.79	25.37	0.26	$<10^{-15}$	$<10^{-15}$	26.24	25.07	25.65	0.26	$<10^{-15}$	$<10^{-15}$
Z	0.998	0.00	0.00	0.00	0.00	>0.999	$4.76\cdot10^{-5}$	0.00	0.00	0.00	0.00	0.993	0.003
N3-Z	$1.10\cdot10^{-14}$	18.96	18.99	18.97	0.01	$1.00\cdot10^{-15}$	$<10^{-15}$	19.19	19.20	19.19	0.00	$2.79\cdot10^{-12}$	$1.20\cdot10^{-14}$
Enol-Z	0.002	3.22	4.56	3.89	0.30	$8.77\cdot10^{-4}$	$4.76\cdot10^{-4}$	2.96	4.27	3.62	0.29	0.007	0.003
rS	>0.999	0.00	0.00	0.00	0.00	>0.999	$2.31\cdot10^{-5}$	0.00	0.00	0.00	0.00	0.999	$2.61\cdot10^{-4}$
Enol-rS	$8.40\cdot10^{-11}$	14.24	13.33	13.79	0.20	$1.47\cdot10^{-11}$	$5.40\cdot10^{-12}$	14.01	13.09	13.55	0.21	$7.00\cdot10^{-9}$	$2.01\cdot10^{-9}$
N3-rS	$1.58\cdot10^{-4}$	5.65	4.87	5.26	0.17	$7.36\cdot10^{-5}$	$2.31\cdot10^{-5}$	5.29	4.53	4.91	0.17	0.001	$2.61\cdot10^{-4}$
dS	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.41\cdot10^{-6}$	0.00	0.00	0.00	0.00	>0.999	$6.24\cdot10^{-5}$
Enol-dS	$5.00\cdot10^{-15}$	20.09	18.79	19.44	0.29	$1.00\cdot10^{-15}$	$<10^{-15}$	20.16	18.86	19.51	0.29	$1.81\cdot10^{-12}$	$7.29\cdot10^{-13}$
N3-dS	$2.40\cdot10^{-5}$	6.85	5.82	6.33	0.23	$1.06\cdot10^{-5}$	$4.41\cdot10^{-6}$	6.70	5.65	6.18	0.23	$1.92\cdot10^{-4}$	$6.24\cdot10^{-5}$

In contrast to the high-pressure results, the reaction free energies of the isoguanine minor tautomers (Table 30) are decreasing with increasing temperature, this leads to increased fractions, which is amplified due to the temperature dependence of the Boltzmann factor. The, very stable, P is only slightly affected by temperature; for the N1-tautomer, no temperature dependent effect on the reaction free energy can be observed. Both of the other tautomers are energetically destabilized with temperature, but their fractions increase slightly due to the Boltzmann factor; even at 363.15 K, nearly only the Watson-Crick tautomer is populated. The P is very tautomer stable. The Z is the other of the problematic Hachimoji bases, the enol-tautomer has a significant fraction. High temperatures additionally decrease the reaction free energy of this tautomer, its fraction increases to 7 ‰ at 363.15 K. The N3-tautomer is energetically destabilized at higher temperatures, but the population increases slightly; overall, this tautomer fraction can be neglected. The enol-tautomer of the dS behaves like the N3-tautomer of the Z, the other minor tautomers of the S, enol- and N3-rS and N3-dS, are stabilized by temperature. This leads to a 1 ‰ N3-rS, respectively a 0.19 ‰ N3-dS fraction at 363.15 K.

As expected, and like for the natural nucleobases, the minor tautomers of the Hachimoji bases show increasing fractions with high temperatures. This is in contrast to the results for high pressures, which stabilize especially the problematic B. High temperatures may lead to increased mutation events in Hachimoji nucleic acids.

Table 31: Results for the backbone variation of purine nucleobases at 278.15 K and 363.15 K (at 1bar). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (278.15 K) and 9th (363.15 K)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (278.15 K)/ 10th (363.15 K)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (278.15 K) and 11-14 (363.15 K). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
7C-A	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.13\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$2.72\cdot10^{-6}$
7C-N1-A	$5.93\cdot10^{-7}$	9.10	7.84	8.47	0.28	$2.22\cdot10^{-7}$	$1.130\cdot10^{-7}$	9.22	7.94	8.58	0.28	$6.87\cdot10^{-6}$	$2.71\cdot10^{-6}$
7C-N3-A	$1.72\cdot10^{-11}$	16.65	12.53	14.59	0.92	$3.44\cdot10^{-12}$	$5.73\cdot10^{-12}$	17.03	12.94	14.99	0.91	$9.57\cdot10^{-10}$	$1.21\cdot10^{-9}$
8N-7C-A	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.64\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$2.69\cdot10^{-6}$
8N-7C-N1-A	$5.12\cdot10^{-6}$	7.27	7.09	7.18	0.04	$2.29\cdot10^{-6}$	$1.64\cdot10^{-7}$	7.33	7.13	7.23	0.04	$4.44\cdot10^{-5}$	$2.69\cdot10^{-6}$
8N-7C-N3-A	$1.43\cdot10^{-10}$	14.58	12.10	13.34	0.55	$3.29\cdot10^{-11}$	$3.30\cdot10^{-11}$	14.85	12.37	13.61	0.56	$6.42\cdot10^{-9}$	$4.94\cdot10^{-9}$
8N-A	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.58\cdot10^{-7}$	0.00	0.00	0.00	0.00	>0.999	$6.02\cdot10^{-6}$
8N-N1-A	$6.67\cdot10^{-6}$	7.18	6.82	7.00	0.08	$3.15\cdot10^{-6}$	$4.58\cdot10^{-7}$	7.28	6.92	7.10	0.08	$5.34\cdot10^{-5}$	$6.01\cdot10^{-6}$
8N-N3-A	$1.21\cdot10^{-10}$	14.83	12.03	13.43	0.63	$2.81\cdot10^{-11}$	$3.19\cdot10^{-11}$	15.15	12.34	13.74	0.63	$5.35\cdot10^{-9}$	$4.66\cdot10^{-9}$
7C-G	>0.999	0.00	0.00	0.00	0.00	>0.999	$1.53\cdot10^{-5}$	0.00	0.00	0.00	0.00	0.999	$8.80\cdot10^{-5}$
7C-Enol-G	$5.58\cdot10^{-5}$	5.82	5.73	5.77	0.02	$2.92\cdot10^{-5}$	$1.03\cdot10^{-6}$	5.69	5.56	5.62	0.03	$4.13\cdot10^{-4}$	$1.62\cdot10^{-5}$
7C-N3-G	$4.92\cdot10^{-5}$	6.44	5.16	5.80	0.29	$2.76\cdot10^{-6}$	$1.43\cdot10^{-5}$	6.86	5.58	6.22	0.29	$1.81\cdot10^{-4}$	$7.19\cdot10^{-5}$
8N-7C-G	>0.999	0.00	0.00	0.00	0.00	>0.999	$7.37\cdot10^{-6}$	0.00	0.00	0.00	0.00	>0.999	$5.12\cdot10^{-5}$
8N-7C-Enol-G	$7.50\cdot10^{-6}$	7.16	6.78	6.97	0.09	$3.36\cdot10^{-6}$	$5.19\cdot10^{-7}$	7.02	6.61	6.81	0.09	$7.94\cdot10^{-5}$	$1.02\cdot10^{-5}$
8N-7C-N3-G	$2.94\cdot10^{-5}$	6.66	5.58	6.12	0.24	$1.56\cdot10^{-5}$	$6.85\cdot10^{-6}$	7.04	5.96	6.50	0.24	$1.22\cdot10^{-4}$	$4.10\cdot10^{-5}$
8N-G	>0.999	0.00	0.00	0.00	0.00	>0.999	$4.08\cdot10^{-6}$	0.00	0.00	0.00	0.00	>0.999	$3.74\cdot10^{-5}$

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
8N-Enol-G	$5.04\cdot10^{-6}$	7.57	6.84	7.20	0.16	$2.19\cdot10^{-6}$	$6.51\cdot10^{-7}$	7.45	6.69	7.07	0.17	$5.56\cdot10^{-5}$	$1.30\cdot10^{-5}$
8N-N3-G	$1.50\cdot10^{-5}$	7.09	5.96	6.53	0.25	$7.46\cdot10^{-6}$	$3.43\cdot10^{-6}$	7.48	6.34	6.91	0.25	$6.93\cdot10^{-5}$	$2.44\cdot10^{-5}$
7C-B	0.660	0.00	0.00	0.00	0.00	0.681	0.019	0.00	0.00	0.00	0.00	0.591	0.022
7C-Enol-B	0.004	1.99	4.19	3.09	0.49	0.003	0.002	1.40	3.59	2.50	0.49	0.019	0.013
7C-N3-B	0.336	0.52	0.32	0.42	0.04	0.316	0.018	0.40	0.20	0.30	0.04	0.391	0.020
8N-7C-B	0.558	0.00	0.00	0.00	0.00	0.578	0.032	0.00	0.00	0.00	0.00	0.498	0.027
8N-7C-Enol-B	0.003	2.43	4.19	3.31	0.39	0.001	0.001	1.87	3.63	2.75	0.39	0.011	0.006
8N-7C-N3-B	0.439	0.33	0.02	0.17	0.07	0.421	0.031	0.17	-0.15	0.01	0.07	0.491	0.027
8N-B	0.787	0.00	0.00	0.00	0.00	0.815	0.048	0.00	0.00	0.00	0.00	0.688	0.061
8N-Enol-B	0.011	1.49	3.82	2.65	0.52	0.007	0.007	0.96	3.30	2.13	0.52	0.036	0.027
8N-N3-B	0.203	0.47	1.21	0.84	0.16	0.178	0.045	0.29	1.03	0.66	0.16	0.276	0.053

The adenine bases with a modified purine backbone (Table 31) show a similar, non-linear behavior like the enol-guanine, but their minor tautomer fractions are smaller at high temperatures. The same temperature dependent behavior is observed for all of the modified guanine species. The tautomer fractions of all of the minor tautomers are in the 10^{-5}-10^{-4} range at 363.15 K, making mispairings more likely if these bases are incorporated in a nucleic acid which is exposed to high temperatures. A non-linear behavior is also observed for the modified isoguanines, here the minor tautomers are energetically destabilized at high and low temperatures, but the fractions of the Watson-Crick tautomers are only increasing at low temperatures since the Boltzmann factor leads to smaller fractions at high temperatures. The isoguanine derivatives behave equally by temperature and pressure variations; the problematic minor tautomer, the N3-tautomer, is energetically destabilized by changes in the environmental conditions, but in the temperature and pressure range investigated in this work, its fraction remains significant and would most likely lead to many mutation events in Hachimoji nucleic acids.

Table 32: Results for the thiolated guanines at 278.15 K and 363.15 K (at 1bar). The investigated tautomers are given in the first column, followed by the ambient condition results. The reaction free energies, ΔG (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM (3rd (278.15 K) and 9th (363.15 K)) and MP2/6-311+G(d,p)/EC-RISM hydration free energy difference/CCSD(T)/cc-pVTZ gas phase reaction free energy (4th (278.15 K)/ 10th (363.15 K)) levels of theory are shown. The average reaction free energies from both EC-RISM approaches, the corresponding error as well as the populations (with errors) are given in columns 5-8 (278.15 K) and 11-14 (363.15 K). The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. The structures and results for the respective conformations are given in the SI part 05.

ΔG / population	Average pop. 298.15 K	EC-RISM 278.15 K	EC-RISM/CCSD(T) 278.15 K	Average ΔG 278.15 K	Error ΔG 1 278.15 K	Average pop. 278.15 K	Error pop. 278.15 K	EC-RISM 363.15 K	EC-RISM/CCSD(T) 363.15 K	Average ΔG 363.15 K	Error ΔG 363.15 K	Average pop. 363.15 K	Error pop. 363.15 K
Thio-B	0.146	0.00	0.00	0.00	0.00	0.146	0.100	0.00	0.00	0.00	0.00	0.132	0.071
Thiol-Thio-B	0.030	-0.10	2.23	1.07	0.52	0.021	0.034	-0.81	1.52	0.35	0.52	0.081	0.093
N3-Thio-B	0.824	-1.95	0.03	-0.96	0.44	0.832	0.128	-2.28	-0.30	-1.29	0.44	0.787	0.145
Thio-G	0.998	0.00	0.00	0.00	0.00	0.999	$1.85\cdot10^{-4}$	0.00	0.00	0.00	0.00	0.994	0.001
Thiol-Thio-G	$2.57\cdot10^{-4}$	6.19	3.62	4.90	0.57	$1.41\cdot10^{-4}$	$1.46\cdot10^{-4}$	6.05	3.52	4.79	0.57	0.001	0.001
N3-Thio-G	0.002	3.84	3.75	3.80	0.02	0.001	$3.86\cdot10^{-5}$	3.92	3.83	3.88	0.02	0.005	$1.36\cdot10^{-4}$

Considering the thiolated species (Table 32), the influence of temperature onto the reaction free energies of thioguanine is negligible, but due to the Boltzmann factor the populations of the minor tautomers increase with temperature. The resulting fractions are 1 ‰ for the thiol- and 5 ‰ for the N3-tautomer at 363.15 K. The thioisoguanine is stronger affected by temperature effects. Both minor tautomers are energetically stabilized at high temperatures, resulting in a N1-tautomer fraction of only 13.2 % at 363.15 K.

The overall tautomer stability of the natural nucleobases is high. This stability is affected by environmental changes like high hydrostatic pressure and temperature variations, but even though more mutation events can be expected at extreme conditions, they are overall tautomer stable due to the high reaction free energies of the minor tautomers at ambient conditions which become only slightly reduced. The Hachimoji nucleobases are overall stabilized by pressure and destabilized by high temperatures. Two of the Hachimoji minor tautomers are problematic because of their large fractions, the B N3- and the Z enol-tautomer. The environmental effects are not destabilizing these tautomers in a way that B and Z are comparably stable to the natural nucleobases. This is also not achieved by thiolation or purine backbone modifications of the isoguanine. The only change from guanine to isoguanine is the switch of the 2-amino- and 6-oxo-groups of guanine. This way, the guanidine motif is destroyed in isoguanine. The stability of this group may play an important role in the tautomer stability and may explain the overall lower tautomer stability of isoguanine.

The natural nucleobases are evolutionary well chosen, they are tautomer stable in a way that mutation events in nucleic acids at early life conditions are supposed to occur more frequently. The Hachimoji nucleic acids, in contrast, are the most stable at high pressures from a tautomeric point of view, which is in contrast to the need of evolution on earth. The increased tautomer stability of the Hachimoji code at high pressures makes this code an interesting system which may play an important role for extraterrestrial life, while the decreased tautomer stability at high-temperatures is problematic on earth, especially at early life conditions. In general, the more unstable nucleobases show a larger variance with temperature, since small energetic changes affect the populations more strongly than for nucleobases with high reaction free energies.

4.6 Nucleotides at high hydrostatic pressure: NMR of adenosine monophosphate

In contrast to small osmolytes like TMAO and nucleobases, nucleotides have much more conformational degrees of freedom. This is partly due to the freedom within the phosphate chain and partly due to the orientation of the phosphate chain and the nucleobase to the ribose. The conformation of the ribose is also important. Before discussing the results, these conformations have to be clarified. As it can be seen in *3.2.6*, the abbreviation for the conformation consists of three parts, the first describing the orientation of the base toward the ribose: *syn* is defined as the base pointing toward the ribose, *anti* away from it (rotation around the N9 to C1' bond in Figure 34 A/B). The second is the puckering of the ribose ring: The N-state is a 3'-*endo* twist, and the S-state a 2'-*endo* twist. The third one is the phosphate chain orientation: the *gt*-conformation is pictured in Figure 34 C, the *tg*- and *gt*-conformations result from a clockwise or counterclockwise 120° rotation of the front part shown in the Newman projection.

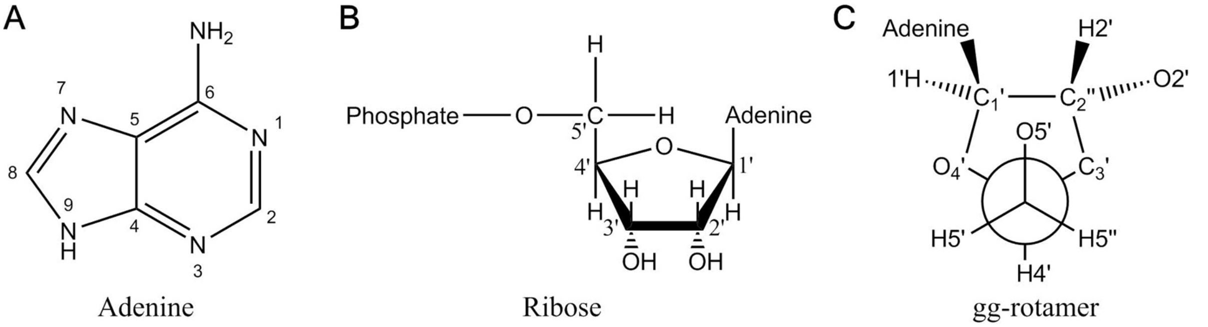


Figure 34: Structure and nomenclature of AMP. Adenine base with atom numbering (A), ribose conformation with atom numbering (B) and rotation of the investigated AMP structure (C); in the latter, a clockwise rotation of 120° of the front part shown in the Newman projection leads to the tg-conformer, counterclockwise rotation to the gt-conformation. Adapted from Ref 6.

Experimentally, all three parts can be obtained by NMR spectroscopy, which was done in Ref 6; the experimental values used here are taken from this reference, the experiments were not performed by the author. To distinguish the *syn*- and *anti*-conformations, the intramolecular distances between the H8 of the adenine and the H1' and H2' of the ribose are used. These distances are measured using T-ROESY (a version of the rotating frame Overhauser effect spectroscopy) cross peaks and calculating the distance using a calibration which amongst others needs the distance between the hydrogen atoms at the 5' position of the ribose *d*(H5″,H5′). The signals of these protons could not be distinguished in Ref. 6 and therefore, no experimental populations are available for AMP. However, the Kalbitzer group did not only measure the mono-, but also the di- and triphosphate (ADP and ATP respectively) as well as the nucleoside, adenosine. For this species, a preference of the *anti*-conformation could be observed with a large fraction of the *syn*-conformation for adenosine (40%) and a strong preference for ADP and ATP (99 and 98%, respectively). The rotamer state of the exocyclic group is also determined using the proton signals of the 5' and 5'' protons, via the ${}^3J$(H5',H4') coupling constant, and a generalized Karplus equation, and therefore, could not be calculated for AMP. In all of the other adenine nucleotides, the *gg*-rotamer was found to be the dominant one with fractions of 75, 78 and 84% for A, ADP and ATP respectively. The *gg*-fraction increases with inclusion and lengthening of the phosphate chain, while the *gt*-fraction is decreasing (21, 15 and 12%, respectively), and the *tg*-fraction remains more or less unchanged with 4, 7 and 4 %. The relative populations of the ribose conformations were not accessible from the T-ROESY measurements due to the indistinguishability of 5' and 5'' in AMP, but again for ADP and ATP, as well as for GDP and GTP. From ROEs, a weak preference of the S-state is expected. Alternatively, the ribose conformation can be predicted from ${}^3J$(H1',H2') coupling constants. Here, too, for AMP and in all cases, the S-state is preferred. While A, AMP and ATP have S-fractions within the same range (77, 74 and 73%, respectively), ADP shows a larger N-fraction with 38%. All these experimental findings can be found in the respective tables in Ref. 6 and are summarized in Table 33.

Table 33: Experimental data for the adenine nucleotides in aqueous solution at ambient conditions. ${}^3J$ coupling constants in Hz, populations in % and interatomic distances in nm are given. Data taken from Ref. 6, see the original publication for experimental details and equations for the calculation of the populations.

	Adenosine	AMP	ADP	ATP
${}^3J$(H5'',H4')	3.4	-	2.9	2.6
${}^3J$(H5',H4')	2.7	-	2.9	2.7
P_{gg}	75	-	78	84
P_{gt}	21	-	15	12
P_{tg}	4	-	7	4
d(H8,H1')	0.26	-	0.36	0.35
d(H8,H2')	0.27	-	0.27	0.26
P_{syn}	40	-	1	2
P_{anti}	60	-	99	98
${}^3J$(H1',H2')	6.32	6.10	5.27	6.01
P_{N}	23	26	38	27
P_{S}	77	74	62	73

The expectations from the experimental data of A, ADP and ATP are that the *gg*-rotamer and the *anti*-conformation should be part of the preferred AMP structure in solution; the analysis of the ${}^3J$(H1′,H2′) coupling additionally predicts a preference for the S-state. The ring puckering is the only conformational estimate directly measured for AMP and therefore an ideal starting point for the comparison of theory and experiment. A relative population P_S of 74% was estimated experimentally, while a value of 83% was obtained computationally (Table 34). This population was calculated considering only two conformations; taking the whole conformational ensemble into account, the results may be improved. This is very close to the NMR estimate considering that the energetic differences related to these populations are within the range of errors which can be expected from EC-RISM using the PMV correction, approx. 1 kcal mol^{-1}. Further considering the uncertainty concerning the AMP conformations underlines the good agreement between theory and experiment. There is a high degree of conformational freedom within the AMP molecule: even neglecting the *gt*- and *tg*-rotamers, which are significantly populated in A, ADP and ATP, there are the ribose-OH conformations, and only two single conformations were used for the theoretical predictions. Further, NMR calculations were used to study the indistinguishability of H5’ and H5’’ chemical shifts in AMP. Thus, the isotropic shielding constants and, using the DSS reference shielding of 31.8997 ppm (see chapter 4.4, respectively Ref. 7), isotropic chemical shifts were calculated, and the difference of H5’ and H5’’ chemical shifts was investigated. Calculated data and experimental chemical shifts are given in Table 34.

Table 34: Calculated data for the two AMP conformations anti-S-gg and anti-N-gg as well as experimental ^1H chemical shifts (in ppm, measured at pH 9.4, data taken from Ref. 6). The table includes: ^1H chemical shifts at ambient pressure (in ppm), δ; intramolecular energies (in kcal mol^{-1}), E_{intra}; excess chemical potential with ($\mu^{ex,corr}$) and without (μ^{ex}) PMV correction (in kcal mol^{-1}); relative total free energy to the minimum free energy structure, calculated via $\Delta G_{rel}=\mu^{ex,corr}+E_{intra}-(\mu^{ex,corr}+E_{intra})_{min}$; PMV (in Å^3); the conformer populations, $x_{conf.}$ and the population weighted calculated isotropic chemical shifts, δ_{calc}. The chemical shifts are relative to a DSS reference shielding constant of 31.8997 ppm.

Structure	*anti*-S-*gg*	*anti*-N-*gg*	δ_{calc}	δ_{exp}
δH2	8.526	8.493	8.521	8.23
δH8	9.296	9.556	9.340	8.62
δH1’	6.519	6.454	6.508	6.13
δH2’	5.348	4.325	5.177	4.80
δH3’	4.578	4.975	4.644	4.50
δH4’	4.695	4.460	4.656	4.36
δH5’	4.082	4.422	4.139	3.99
δH5”	3.952	4.152	3.986	3.99
E_{intra}	-958071.07	-958055.22		
μ_{ex}	-301.49	-316.40		
$\mu_{\text{ex,corr}}$	-289.65	-304.56		
G_{rel}	0	0.95		
PMV	183.854	183.807		
x_{conf}	0.833	0.167		

The experimental H2 chemical shift signal is upfield shifted to the H8 resonance by 0.39 ppm. A similar behavior is observed for the other adenine nucleotides,[6] which is in agreement with theory that predicts an upfield shift of 0.93 ppm. The comparison of these two chemical shifts serves as a benchmark for the quality of the chemical shift prediction using only two structures since both protons investigated are bound to the base and therefore independent of ribose OH-rotations. Because this benchmark confirms the computational setup as reasonable, even using only two conformations, the H5' and H5'' chemical shifts can be investigated. Experimentally indistinguishable with a chemical shift of 3.99 ppm, these shifts are unique within the adenine nucleotides, having an upfield shift of the H5'' resonance compared to the H5' signal for adenosine (-0.08 ppm), and a downfield shift for ADP (+0.08 ppm) and ATP (+0.09 ppm).[6] This equality of chemical shifts is not obtained from the calculation. Here, a more adenosine-like situation is observed, with a downfield shift of the H5'' resonance of -0.15 ppm using the conformational averaged shielding constant and of -0.13 and -0.27 for the S- and N-state, respectively. The absolute value of the H5'/H5'' is larger than in the other nucleotides, leading to the conclusion that an internal rotation of this part of the molecule is not a sufficient explanation for the experimental indistinguishability of these shifts. A possible explanation for this phenomenon is the chemical shift's anisotropy (CSA) related to the phosphorus, which is not captured by the calculation due to the fact that only isotropic shielding constants are computed, detailed discussions of this effect are in chapter 4.7.3. A downfield shift of H5'' resonance compared to the H5' would explain the trend of experimental differences between these two nuclei and would lead, together with the inequality of the resonances obtained from theory, to the conclusion that the indistinguishability of AMP H5' and H5'' signals is an unfortunate coincidence of chemical shift anisotropy resulting from the addition of the phosphate chain to adenosine resulting in a cancellation of the signal difference.

To investigate the pressure dependence of AMP chemical shifts, calculations and experiments were done in the range from 1 bar to 2 kbar (using the DSS chemical reference shielding constants described in chapter 4.4). The pressure coefficients from a fit of the experimental and calculated data to a second order polynomial are given in Table 35. The experimental AMP pressure dependence is well represented for the H2 and H4' nuclei by both pressure coefficients having the same sign, but the pressure dependence is slightly underestimated. The experimental H1' first order coefficient is negative but near zero, while all calculated ones are strictly positive. Here, the pressure dependence is only reflected by the sign of the second order coefficient. In case of the H8, H2’, H3’ and the experimentally indistinguishable H5’ and H5’’, the calculated pressure coefficients do not match the sign of the experimental ones. The calculations are not able to reproduce the pressure dependence correctly in this case. The calculation shows a stronger pressure dependence of the H5’’compared to H5’, which should result in a distinguishability of both signals upon pressurization, which is not found experimentally. The pressure dependence of the AMP is here investigated by using only two structures of which it is known experimentally that they are not representative for the whole conformational ensemble of AMP. Besides, hydroxyl-rotations are neglected here. Therefore, the pressure dependence of the conformational ensemble of AMP, which is measured by the NMR experiments, cannot be reproduced by using only two structures. A proper conformational sampling and calculations on the majority of the ensemble seem to be necessary to correctly reproduce the experiments, which comes with a great increase of computational cost, especially when NMR parameters have to be calculated for every conformation. An approach for the calculation of NMR chemical shifts using MD data and explicit solvation is described in S. Maste’s master’s thesis (Ref. 297).

Table 35: Linear (B_1, in ppm GPa^{-1}) and quadratic (B_2, in ppm GPa^{-2}) coefficients from fitting experimental AMP, ADP and ATP and calculated population weighted AMP chemical shifts and the ones of the single conformations to the following form: $\delta(p)=\delta_0+B_1p+B_2p^2$.

Structure	H2	H8	H1‘	H2‘	H3‘	H4‘	H5‘	H5‘‘
ATP(exp.)								
B_1	0.134	-0.113	-0.001	-0.099	-0.083	0.092	0.083	0.019
B_2	-0.220	0.168	-0.020	0.151	0.032	-0.060	-0.070	-0.102
ADP(exp.)								
B_1	0.134	-0.090	0.005	0.029	-0.164	0.113	0.023	-0.100
B_2	-0.210	0.147	-0.055	0.006	0.210	-0.040	-0.040	0.100
AMP(exp.)								
B_1	0.118	-0.131	-0.001	-0.084	-0.026	0.104	-0.029	-0.029
B_2	-0.200	0.190	-0.056	0.048	0.040	-0.120	0.056	0.056
AMP(calc.)								
B_1	0.063	0.034	0.054	0.045	0.098	0.073	0.060	0.075
B_2	-0.045	-0.022	-0.029	-0.008	-0.082	-0.051	-0.057	-0.060
anti-S-gg								
B_1	0.063	0.028	0.052	0.067	0.097	0.077	0.050	0.071
B_2	-0.046	-0.010	-0.029	-0.048	-0.072	-0.059	-0.043	-0.053
anti-N-gg								
B_1	0.067	0.016	0.076	0.118	0.032	0.096	0.053	0.060
B_2	-0.049	-0.012	-0.048	-0.083	-0.026	-0.074	-0.041	-0.043

4.7 Nucleotides at high temperatures

There are several studies regarding the tautomerism of free nucleobases. Thereby, natural[300,301,302,303,304] and non-natural[306,307,308,312,313] nucleobases were investigated, theoretically and experimentally, in the gas-phase as well as in free solution. Rare tautomers of natural nucleobases could be identified and stabilized in the presence of metal ions,[314,315,316] or in charged complexes,[317] and there is evidence that natural[318] and non-natural[305,307,308,318,319] nucleobase rare tautomers lead to mispairings in nucleotide duplexes and strands (rare tautomer hypothesis). Even though this hypothesis is questioned in the literature[320] there is evidence that the rare tautomers lead to mispairings and the formation of Wobble-pairs.[321] Modern crystal structure data suggests that the A-C mispairing, which is possible when alternative tautomers are involved, is stabilized by the DNA polymerase.[322] But there is less work regarding the tautomers of nucleosides and nucleotides in free solution. NMR experiments indicate that the pure GMP and CMP are tautomer stable[323] while newer data suggests that the tautomerism of nucleobases in strands is sequence dependent and leads to G-T/U mispairings with involvement of G- and T- respectively U-enol-tautomers. These G-, T-, and U-enol-tautomers are supposed to have fractions of less than 0.4 %, in line with own computational data discussed in chapter 4.5.[324] The situation in an nucleic acid strand seems to be extremely complicated and not only tautomer dependent, also the isostericity and many different geometric aspects play an important role in this environment.[325] Since the situation in a nucleic acid strand is to complicated and the system size is too large to study it with help of QC calculations sufficiently, this work focuses on the free nucleosides and nucleotides in solution. For these systems, very few literature data is available since the most authors focus on the nucleobases themselves like it is done in chapter 4.5 of this work. Especially the detailed analysis of the tautomeric behavior of these species using a proper conformational sampling and the investigations of the temperature dependence is a new aspect which can be helpful for other researchers to close the gap between the situation in nucleobases and larger nucleic acids.

4.7.1 Experimental results

In chapter 4.6, the main conformations of AMP were studied with the help of EC-RISM calculations and experimental ^{1}H chemical shifts. The restrictions limiting this analysis are that only two conformations are considered and that only the ^{1}H chemical shifts are experimentally available. However, the histamine analysis in chapter 4.3 revealed that the ^{15}N chemical shifts are the most important for the investigation of prototopic tautomerism taking place at nitrogen atoms. The nucleoside and nucleotide analysis presented here is taking multiple conformations into account (the computational details are described in chapter *3.2.7*), also the chemical shifts of all ^{15}N, ^{13}C and ^{1}H nuclei of the nucleobases are measured (done by the author in Roland Sigel's laboratory at UZH Zürich during an internship funded by RESOLV). The experimental chemical shifts of the desoxyribonucleosides and ribonucleotides are given in Table 36 to Table 39 (for the calculations the same sugars are used as in the experiment. The calculations on the thymidine-triphosphate, which is not used in the experiment, are done using the desoxyribose). The ^{1}H sugar resonances are not given since they interfere with the water and water suppression signals and therefore could not be analyzed correctly. Some of the resonances of the exchanging protons - the H1 and H2 from the guanine bases, the H6 from adenine bases, the H3 from thymine bases, and the H4 from cytosine bases (with separate signals for both protons) - are visible in the NMR at low temperatures, which can be helpful by determining tautomer fractions via NMR. The assignment of the experimental chemical shifts to respective nuclei of the nucleotides, is done with help of the Biological Magnetic Resonance Data Bank (BMRB, https://bmrb.io/ref_info/stats.php?restype=dna&set=filt) as well as ^{1}H/^{13}C-HSQC experiments.

Table 36: Experimental ^{1}H chemical shifts of the bases of the nucleosides and nucleotides. Obtained by direct ^{1}H measurements of the compounds at the respective temperatures given. The atom numbering is similar as for the nucleobases (Figure 29).

Compound				Compound			
Adenosine	H2	H6	H8	ATP	H2	H6	H8
278.15	8.2596	6.8163	8.1287	278.15	8.5302	6.8701	8.2265
298.15	8.2701	6.7699	8.1744	298.15	8.5136	6.7903	8.2351
323.15	8.2690	6.6826	8.2067	323.15	8.4853	6.6888	8.2405
348.15	-	-	-	348.15	8.3979	-	8.1803
Guanosine	H1	H2	H8	GTP	H1	H2	H8
278.15	11.1007	6.3730	7.9593	278.15	-	6.3928	8.1314
298.15	-	-	7.9512	298.15	-	6.3250	8.1047
323.15	-	-	7.9340	323.15	-	-	8.0707
348.15	-	-	7.8524	348.15	-	-	7.9787
Thymidine	H3	H6	H7	UTP	H3	H5	H6
278.15	11.2155	7.6459	2.3534	278.15	-	5.3491	7.9576
298.15	11.0396	7.5999	2.3290	298.15	-	5.3190	7.9330
323.15	11.0831	7.5818	2.3421	323.15	-	5.3100	7.9250
348.15	-	7.4627	2.2739	348.15	-	5.2585	7.8676
Cytidine	H4	H5	H6	CTP	H4$_1$/H4$_2$	H5	H6
278.15	7.1672	6.0552	7.8652	278.15	6.7476/7.3303	5.9804	7.9831
298.15	-	6.0125	7.7945	298.15	-	5.9743	7.9511
323.15	-	6.0145	7.7700	323.15	-	5.9665	7.9193
348.15	-	5.4671	7.2483	348.15	-	5.8922	7.8303

Table 37: Experimental ^{15}N chemical shifts of the bases of the nucleosides and nucleotides. Obtained by direct ^{15}N measurements of the compounds at the respective temperatures given. The atom numbering is similar as for the nucleobases (Figure 29).

Compound									
Adenosine	N1	N3	N6	N7	N9	Thymidine	N1	N3	
278.15	224.6530	216.1690	78.6140	229.6280	172.6040	278.15	157.0800	147.5510	
298.15	225.2520	216.9580	78.2520	229.8840	172.9880	298.15	156.9310	147.4380	
323.15	226.2040	218.0060	77.6350	230.3420	173.4290	323.15	156.7890	147.2450	
348.15	226.8150	218.6800	76.8260	230.5910	173.5720	348.15	156.6070	147.1030	
Guanosine	N1	N2	N3	N7	N9	Cytidine	N1	N3	N4
278.15	146.9490	72.5410	165.1640	234.2430	172.9740	278.15	156.6500	190.5890	95.7130
298.15	146.9560	72.3620	165.5590	234.5320	173.2180	298.15	156.7150	195.0200	93.7690
323.15	147.0820	72.0330	166.1040	235.1070	173.6980	323.15	156.9250	199.1150	92.9350
348.15	147.0960	71.2970	166.4970	235.5660	173.8340	348.15	156.1890	194.5890	93.5730
ATP	N1	N3	N6	N7	N9	UTP	N1	N3	
278.15	225.5120	216.0840	78.3190	231.9470	168.8220	278.15	146.0970	158.6670	
298.15	226.1600	216.7120	77.9700	232.0590	168.8110	298.15	145.8610	158.7140	
323.15	226.8610	217.6690	75.3700	232.2400	168.7230	323.15	145.6510	158.9140	
348.15	227.4190	218.4640	76.8640	232.4220	169.1040	348.15	145.3480	158.7900	
GTP	N1	N2	N3	N7	N9	CTP	N1	N3	N4
278.15	165.9800	72.9170	147.2960	235.6670	168.3100	278.15	152.2840	198.7720	92.1600
298.15	166.3820	72.6700	147.5000	235.8560	168.3920	298.15	152.4640	200.1760	93.2800
323.15	166.9220	72.3440	147.8940	236.1990	168.4920	323.15	152.5330	201.4590	92.6700
348.15	167.2730	72.1470	148.1850	236.5660	169.7430	348.15	152.6110	202.3540	92.0800

Table 38: Experimental ^{13}C chemical shifts of the bases of the nucleosides. Obtained by direct ^{13}C measurements of the compounds at the respective temperatures given. The atom numbering is similar as for the nucleobases (Figure 34 and Figure 29).

Compound										
Adenosine	C2	C4	C5	C6	C8	C1‘	C2‘	C3‘	C4‘	C5‘
278.15	154.8547	150.8100	121.5670	158.1663	142.9205	90.2637	41.7238	74.1364	87.4520	64.4645
298.15	155.0798	151.0430	121.7150	158.3259	143.0478	90.2375	41.7594	74.0974	87.4096	64.5098
323.15	155.3239	151.2730	121.8670	158.4785	143.1609	90.2330	41.8377	74.0797	87.3739	64.5900
348.15	155.4089	151.3780	121.9070	158.5112	143.1654	90.1601	41.8475	74.0299	87.2992	64.6111
Guanosine	C2	C4	C5	C6	C8	C1‘	C2‘	C3‘	C4‘	C5‘
278.15	156.5375	153.7899	119.0090	161.6740	140.3404	89.9055	41.3751	73.9932	86.6140	64.3726
298.15	156.5692	153.8344	119.1330	161.6820	140.4090	89.9036	41.4311	73.9744	86.6218	64.4411
323.15	156.5948	153.8809	119.2940	161.6570	140.4683	89.9283	41.5301	73.9916	86.6747	64.5560
348.15	156.5576	153.8542	119.3640	161.5570	140.4216	89.8988	41.5617	73.9712	86.6629	64.5963
Thymidine	C2	C4	C5	C6	C7	C1‘	C2‘	C3‘	C4‘	C5‘
278.15	154.5033	169.4117	114.1070	140.1896	14.3335	89.1730	41.1744	73.1402	87.5743	63.8573
298.15	154.4488	169.3041	114.0990	140.1662	14.2658	89.1948	41.1744	73.1654	87.6619	63.9348
323.15	154.4114	169.1650	114.1040	140.1724	14.1973	89.2977	41.2468	73.2803	87.8584	64.1168
348.15	154.3100	168.9692	114.0210	140.0927	14.0578	89.3109	41.2468	73.3099	87.9320	64.1815
Cytidine	C2	C4	C5	C6		C1‘	C2‘	C3‘	C4‘	C5‘
278.15	158.6010	167.6270	98.7700	144.6588		89.4315	41.9315	73.1942	88.7979	63.9191
298.15	159.1680	168.1420	98.8432	144.5534		89.4238	41.9445	73.2653	88.8261	64.0415
323.15	159.6360	168.6250	98.8492	144.4918		89.4357	41.9752	73.3640	88.8795	64.1995
348.15	158.6200	167.5920	98.2729	143.9720		88.8509	41.3719	72.6950	88.2528	63.4723

Table 39: Experimental ^{13}C chemical shifts of the bases of the nucleotides. Obtained by direct ^{13}C measurements of the compounds at the respective temperatures given. There is a small alternative C5' chemical shift for rATP and rCTP, indicating the existence of an alternative sugar conformation. The atom numbering is similar as for the nucleobases (Figure 34 and Figure 29).

Compound										
ATP	C2	C4	C5	C6	C8	C1‘	C2‘	C3‘	C4‘	C5‘/C5‘$_2$
278.15	155.4241	151.7551	121.2540	158.3603	142.4596	89.1640	76.9712	73.1183	86.7825	67.8003/ 67.1802
298.15	155.5681	151.8638	121.3510	158.4391	142.6203	89.2935	77.0050	73.0960	86.7814	67.8301/ 67.2005
323.15	155.7081	151.9985	121.4830	158.5136	142.8081	89.5145	77.0570	73.1074	86.7798	67.8942/ 67.3272
348.15	155.7390	152.0407	121.5310	158.5136	142.8793	89.6498	77.0372	73.0814	86.7222	67.8950/ 67.3769
GTP	C2	C4	C5	C6	C8	C1‘	C2‘	C3‘	C4‘	C5‘
278.15	156.8108	154.5534	118.7690	161.8830	140.3374	89.1292	76.3038	73.0617	86.6544	67.7900
298.15	156.8199	154.5526	118.9000	161.8638	140.4511	89.2875	76.3136	73.0720	86.6641	67.8520
323.15	156.8472	154.5461	119.0760	161.8596	140.5664	89.5422	76.3509	73.1069	86.6663	67.9310
348.15	156.8208	154.4759	119.1600	161.8091	140.5644	89.7040	76.3352	73.0972	86.6090	67938
UTP	C2	C4	C5	C6		C1‘	C2‘	C3‘	C4‘	C5‘
278.15	154.6553	169.1995	105.2267	144.3956		90.8649	72.1394	76.5700	85.9440	67.4228
298.15	154.6627	169.1254	105.3082	144.3938		90.8927	72.2388	76.5108	86.0110	67.5016
323.15	154.7108	169.0520	105.4062	144.4272		91.0382	72.3843	76.4783	86.1080	67.6173
348.15	154.7085	168.9452	105.4123	144.4028		91.1741	72.4566	76.4313	86.1320	67.6597
CTP	C2	C4	C5	C6		C1‘	C2‘	C3‘	C4‘	C5‘/C5‘$_2$
278.15	160.1808	168.6833	91.6982	144.1168		99.3760	71.8196	77.0029	85.3952	67.2292/ 66.6920
298.15	160.3180	168.8634	91.7492	144.1878		99.4090	71.9676	77.0242	85.4744	67.3394/ 66.8389
323.15	160.4065	169.0404	91.8656	144.3166		99.3940	72.1423	77.0683	85.5769	67.4586/ 67.0047
348.15	160.3728	169.1015	91.9440	144.3624		99.2890	72.2177	77.0546	85.6002	67.4839/ -

The temperature dependent trends of the ^{1}H chemical shifts (Table 36) are monotonous with temperature in most cases. This indicates that there are no major switches in the population of the main tautomers, since the calculated, temperature dependent, shielding constants (presented in chapter 4.4) are also monotonous with temperature. The ^{15}N chemical shifts (Table 37), the most useful nuclei for tautomer predictions from NMR (since the tautomerism takes place, like for histamine, at the nitrogen atoms) are resolved for each species at each temperature, allowing the temperature dependent NMR fitting for all of the investigated species. While, like for the ^{1}H resonances, the temperature dependent trends are mostly monotonous, there are some exceptions. These are the N6 chemical shifts of ATP at 348.15 K and all of the shifts of cytidine, also at 348.15 K, and some of the CTP resonances at different temperatures. This leads to the assumption that the cytosine-containing compounds are most likely to undergo some major tautomer shifts. The ^{13}C chemical shifts (Table 38 and Table 39) are again mostly monotonous in temperature trends, they could be distinguished completely for all species, including the sugar carbons, since they, unlike the ^{1}H shifts, are not able to interfere with the water signal. Here two different, separated resonances are observed for the C5' of ATP and CTP. Although being very small, these resonances indicate the existence of a minor species with a small but significant and stable fraction.

4.7.2 Nucleosides

The tautomer populations, from EC-RISM energy calculations, were obtained for each of the temperatures given in the computational details. Since the changes in tautomer fractions are monotonous, the results at the highest and lowest temperature used for the parametrization of the EC-RISM^{T} correction are explicitly presented and discussed here, all other results as well as the structures and raw data can be found in SI part 07. The results for the nucleosides are given in Table 40. The data presented are the Boltzmann-weighted energies over all conformations (sampled via MD simulation as described in chapter *3.2.7*) of the respective tautomers used for EC-RISM calculations and the populations calculated using these energies.

Table 40: Results for the nucleosides at 278.15 and 363.15 K. The investigated tautomers are given in the first column; the reaction free energies $\Delta_r G$ (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM and MP2/6-311+G(d,p)/EC-RISM^{T} levels of theory are shown in columns 2-5. The corresponding populations are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer.

ΔG /	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	x_{taut}	x_{taut}	x_{taut}	x_{taut}
population	EC-RISM	EC-RISM^{T}	EC-RISM	EC-RISM^{T}	EC-RISM	EC-RISM^{T}	EC-RISM	EC-RISM^{T}
	278.15 K	278.15 K	363.15 K	363.15 K	278.15 K	278.15 K	363.15 K	363.15 K
Adenosine	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
N1-Adenosine	11.20	11.19	8.75	8.76	$1.60\cdot10^{-9}$	$1.61\cdot10^{-9}$	$5.4\cdot10^{-6}$	$5.4\cdot10^{-6}$
N3-Adenosine	17.80	17.78	13.82	13.86	$1.03\cdot10^{-14}$	$1.07\cdot10^{-14}$	$4.8\cdot10^{-9}$	$4.6\cdot10^{-9}$
Guanosine	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
Enol-Guanosine	6.64	6.65	5.08	5.05	$6.03\cdot10^{-6}$	$5.94\cdot10^{-6}$	$8.8\cdot10^{-4}$	$9.1\cdot10^{-4}$
N3-Guanosine	6.79	6.78	5.41	5.42	$4.63\cdot10^{-6}$	$4.70\cdot10^{-6}$	$5.5\cdot10^{-4}$	$5.4\cdot10^{-4}$
Cytidine	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
Enol-Cytidine	22.27	22.24	17.11	17.14	$3.20\cdot10^{-18}$	$3.33\cdot10^{-18}$	$5.1\cdot10^{-11}$	$4.9\cdot10^{-11}$
N3-Cytidine	7.30	7.29	5.46	5.48	$1.84\cdot10^{-6}$	$1.88\cdot10^{-6}$	$5.2\cdot10^{-4}$	$5.0\cdot10^{-4}$
Thymidine	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
2-Enol-Thymidine	13.88	13.86	10.88	10.92	$1.25\cdot10^{-11}$	$1.29\cdot10^{-11}$	$2.8\cdot10^{-7}$	$2.7\cdot10^{-7}$
4-Enol-Thymidine	10.07	10.07	7.89	7.90	$1.21\cdot10^{-8}$	$1.23\cdot10^{-8}$	$1.8\cdot10^{-5}$	$1.8\cdot10^{-5}$

ΔG /	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	x_{taut}	x_{taut}	x_{taut}	x_{taut}
population	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T
	278.15 K	278.15 K	363.15 K	363.15 K	278.15 K	278.15 K	363.15 K	363.15 K
Uridine	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
2-Enol-Uridine	14.38	14.37	11.17	11.20	$5.00 \cdot 10^{-12}$	$5.14 \cdot 10^{-12}$	$1.9 \cdot 10^{-7}$	$1.8 \cdot 10^{-7}$
4-Enol-Uridine	9.05	9.05	7.04	7.03	$7.76 \cdot 10^{-8}$	$7.76 \cdot 10^{-8}$	$5.8 \cdot 10^{-5}$	$5.9 \cdot 10^{-5}$

All natural nucleosides are tautomer stable in the temperature range of 278–363 K. The energetic differences between EC-RISM and EC-RISM^T are very small, so no differentiation between both approaches has to be made in the discussion of the results. In all cases, the Watson-Crick tautomer is the dominant species at low and high temperatures, but the minor tautomers are slightly less penalized at higher temperatures. This is in agreement with the experimental NMR chemical shifts, which show monotonous trends which are expected to be the inherent temperature dependence of the chemical shifts. As expected, the energy differences are largest for the least populated tautomers, but the energetic ordering of the tautomers is unchanged. The highest populated minor tautomer is the enol-guanosine at 363 K with a fraction of approximately 0.00091; also, the N3-guanosine and the N3-cytidine are significantly populated at 363 K with fractions in the 10^{-4} range. Like the nucleobases, the nucleosides are tautomer stable over a broad range of temperatures. This shows that the sugar has no significant influence on the tautomer stability and the nucleosides behave more like the pure nucleobases in free solution. More influence on the tautomeric behavior can be expected from the phosphorylated species, since in this work only the situation in free, infinite diluted solution is studied. In this environment, the nucleic acid building blocks are conformationaly not as much restricted as in a nucleic acid, and therefore have the possibility to fold, which may stabilize rare tautomers.

Before obtaining the tautomer fractions from the NMR fit, as benchmarked for histamine, the deviation of the calculated NMR chemical shifts, calculated for the single tautomers, from the experimental chemical shifts measured at the respective temperatures is analyzed. The error metrics (RMSE, MAE and MSE) and linear regression parameter (slope m, intercept b and coefficient of determination R^2) for each of the nucleoside tautomers at each temperature for which experimental NMR chemical shifts are available are given in Table 41, the data are depicted in Figure 35. To calculate these data, all types of chemical shifts (^{1}H, ^{13}C and ^{15}N) of the base of the nucleosides were used, the ^{13}C chemical shifts of the sugar is not included. All chemical shifts were calculated using the direct temperature dependent referencing, the raw shielding constants are given in the NMR subfolder of SI part 07. But in principle, all of the different referencing methods introduced in chapter 4.4 can be applied to these shielding constants.

Table 41: Error metrics (RMSE, MSE and MAE in ppm) of the nucleoside base NMR ^{1}H, ^{13}C and ^{15}N chemical shifts calculated for the tautomers at the different temperatures (using the tautomer fractions given in Table 40, the full data are in the SI part 07) to the respective corresponding experimental data (presented in Table 36 to Table 39; calculations and experiments performed at 278, 298, 323 and 348 K) as well as the linear regression data slope m, intercept b in ppm and coefficient of determination (R^2) are shown.

Error	RMSE	MAE	MSE	*m*	*b*	R^2
	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K
	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)
Adenosine	10.69/10.51	7.238/7.152	-5.785/-5.626	1.031/1.030	1.442/1.424	0.986/0.987
	10.28/-	7.048/-	-5.424/-	1.028/-	1.519/-	0.987/-
N1-Adenosine	27.48/27.89	15.60/15.82	-4.789/-4.640	0.874/0.870	22.35/22.75	0.858/0.854
	28.44/-	16.10/-	-4.440/-	0.864/-	23.35/-	0.849/-
N3-Adenosine	33.74/34.21	20.09/20.32	-5.620/-5.494	0.871/0.868	23.49/23.87	0.792/0.787
	34.84/-	20.59/-	-5.333/-	0.863/-	24.46/-	0.781/-
Guanosine	12.20/12.04	8.762/8.632	-7.201/-7.092	1.007/1.007	6.177/6.118	0.970/0.971
	11.80/-	8.464/-	-6.913/-	1.006/-	6.139/-	0.972/-
Enol-Guanosine	25.83/25.83	16.17/16.09	-14.04/-13.99	1.017/1.017	11.66/11.66	0.872/0.872
	25.79/-	15.98/-	-13.89/-	1.015/-	11.74/-	0.872/-
N3-Guanosine	24.86/25.98	15.80/15.80	-7.132/-7.035	0.981/0.980	9.752/9.777	0.840/0.839
	25.10/-	15.83/-	-6.870/-	0.978/-	9.884/-	0.837/-
Cytidine	11.47/10.74	7.815/7.407	-7.794/-7.405	1.085/1.073	-1.857/-0.954	0.991/0.992
	10.08/11.23	6.910/7.872	-6.910/-7.872	1.063/1.075	-0.320/-0.687	0.992/0.992
Enol-Cytidine	28.14/28.81	14.52/14.22	-13.61/-13.29	1.001/0.985	13.48/14.98	0.872/0.862
	29.37/29.90	13.87/14.68	-12.88/-13.94	0.973/0.985	15.96/15.63	0.854/0.854
N3-Cytidine	28.38/29.82	14.79/15.74	-7.039/-6.658	0.902/0.883	18.19/20.08	0.824/0.806
	31.01/30.27	16.54/15.76	-6.175/-6.943	0.866/0.879	21.56/20.75	0.792/0.801
Thymidine	12.04/12.03	7.483/7.465	-7.407/-7.393	1.066/1.065	0.791/0.824	0.986/0.986
	12.02/12.00	7.425/7.422	-7.350/-7.361	1.065/1.064	0.831/0.927	0.986/0.986
2-Enol-Thymidine	26.91/27.01	14.30/14.32	-14.23/-14.26	1.130/1.130	1.153/1.195	0.928/0.927
	27.14/27.26	14.33/14.37	-14.26/-14.32	1.130/1.129	1.221/1.332	0.926/0.925
4-Enol-Thymidine	31.57/31.70	16.83/16.91	-16.67/-16.73	1.174/1.174	-0.839/-0.796	0.914/0.913
	31.85/31.99	16.99/17.08	-16.77/-16.86	1.174/1.174	-0.773/-0.665	0.912/0.911

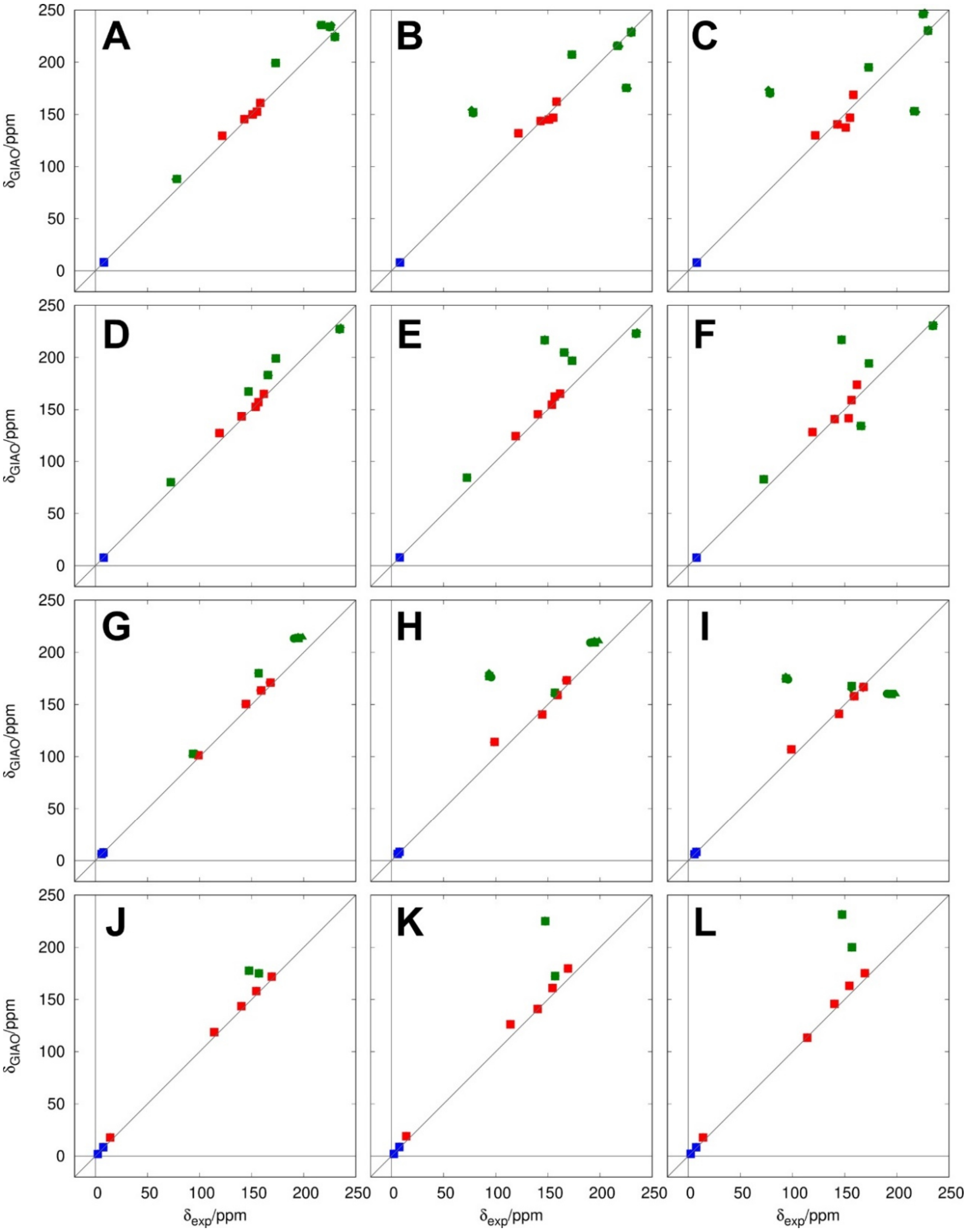


Figure 35: Calculated chemical shifts of the nucleoside tautomers and the experimental results. Shown are all nuclei of the respective bases (hydrogen atoms in blue, carbon atoms in red and nitrogen atoms in green), at 278 (points), 298 (squares), 323 (triangles) and 348 K (rhombs), the temperature dependence is only slightly visible, the error metrics and linear regression parameters are given in Table 41. A-C: Adenosine Watson-Crick-, N1-, and N3-tautomer. D-F: Guanosine Watson-Crick-, enol-, and N3-tautomer. G-I: Cytidine Watson-Crick-, enol-, and N3-tautomer. J-L: Thymidine Watson-Crick-, 2-enol-, and 4-enol-tautomer.

For all nucleosides, the RMSEs of the Watson-Crick tautomers are the smallest, regardless of the temperature and, in most cases, declining with an increasing temperature. Especially for the nitrogen chemical shifts calculated for the minor tautomers a large deviation to the experimental values is observed. The exception is cytidine at the highest temperature measured (348 K), for which the RMSE is the highest; at all other temperatures, the cytidine RMSE follows the general trend observed. Since the errors are calculated between the calculated chemical shifts of the tautomers and the experimental ones of the ionization states, this hints that the Watson-Crick tautomers are the dominant species. The same trends are observed for the behavior of the Watson-Crick tautomer MAEs, including the outlier for cytidine at 348 K. The MSEs have the lowest absolute values of the error metrics, and all have a negative sign. In most cases they show monotonous trends with increasing temperature. In contrast to the RMSEs and MAEs, the Watson-Crick tautomers are not special here by showing the (by far) lowest values. Instead, for most nucleosides (thymidine is the exception), multiple tautomers show values in the same range. The correlation between calculated and experimental chemical shifts, by calculating unsigned error metrics, indicates that the Watson-Crick tautomers are the dominant species of the nucleosides, especially at high temperatures. The linear regression between calculation and experiment results in the largest coefficients of determination for the Watson-Crick tautomers. The coefficient is, as expected from the error metrics, increasing at higher temperatures. The linear regression parameters, too, support the dominance of the Watson-Crick tautomers: the slopes of them are the closest to one, with exception of the cytidine (here the slopes of the enol tautomer are in a similar range), and the intercepts have the smallest absolute values. The intercepts of many of the minor tautomers are very high, indicating a strong difference between computed and experimental chemical shifts for these tautomers and indicate that these tautomers are not populated. The error metrics and linear regression parameters indicate a strong preference for the Watson-Crick tautomers of the nucleosides, which is in agreement with the energy calculations. To quantify this, the populations of the tautomers can be obtained from the experimental NMR chemical shifts, as it is done for histamine, by fitting the calculated chemical shifts of the tautomers to the experimental shifts of the ensemble.

While the errors and linear regressions were calculated using all chemical shifts of the nucleoside base, the fitting of the tautomer populations was done by considering the ^{1}H and ^{15}N chemical shifts only. This is according to the benchmarking done for the histamine. Besides, only the norm$_{\mathrm{N}}$ normalization of the chemical shift values was applied here. The results of the fitting process are given in Table 42 (misleading are the error metrics presented there, these values can, with the mathematica implementation used for the fitting, only be calculated under the assumption of an unconstrained model. The fitting procedure introduced in chapter 3.2.3.2 in contrast, involves constraints: no negative values are allowed for tautomer fractions and the sum of all fractions needs to be exactly 1. Therefore, these errors are not fully reliable but still indicate a quite large uncertainty range). The Watson-Crick tautomers are, again, the dominant species, but their fractions are slightly smaller than the energy calculations suggest. Also, they are, as can be expected from the error metrics, stabilized by increasing temperature (again, the cytidine at 348 K is the exception). The highest populated minor tautomer is the enol-cytidine with a fraction of 0.087 at 278 K. The guanosine is the most stable species with the strongest populated minor tautomer (the N3-tautomer) having a fraction of $1.5 \cdot 10^{-4}$ at 323 K. The errors from the fitting process presented in Table 42 are high, especially for the thymidine. Therefore, the EC-RISM calculations and error metrics from the NMR are more reliable. But the NMR fit is still helpful by investigating the order of magnitude of the tautomer fractions as well as the temperature dependent trends.

Table 42: Results of the calculation of nucleoside tautomer fractions from fitting calculated NMR chemical shifts of the tautomers (given in the NMR subfolder of SI part 07) to the experimental chemical shifts of the base (Table 36 to Table 39) at various temperatures (x_{taut}) together with the errors and the EC-RISM results. The results were obtained by using the ^{1}H and ^{15}N chemical shifts only. The chemical shifts were normalized applying Eqn. 197.

Population	x_{taut}(278 K)	Error(278 K)	EC-RISM(278 K)	x_{taut}(298 K)	Error(298 K)	EC-RISM(298 K)	x_{taut}(323 K)	Error(323 K)	EC-RISM(323 K)	x_{taut}(348 K)	Error(348 K)	EC-RISM(348 K)
A	0.976	0.158	>0.999	0.982	0.153	>0.999	0.992	0.149	>0.999	-	-	>0.999
N1-A	$6.0 \cdot 10^{-6}$	0.193	$1.61 \cdot 10^{-9}$	$7.0 \cdot 10^{-6}$	0.188	$2.05 \cdot 10^{-8}$	$5.4 \cdot 10^{-6}$	0.184	$2.59 \cdot 10^{-7}$	-	-	$1.94 \cdot 10^{-6}$
N3-A	0.024	0.146	$1.07 \cdot 10^{-14}$	0.018	0.143	$6.39 \cdot 10^{-13}$	0.008	0.141	$3.69 \cdot 10^{-11}$	-	-	$9.10 \cdot 10^{-10}$
G	>0.999	0.423	>0.999	>0.999	0.413	>0.999	>0.999	0.402	>0.999	-	-	>0.999
Enol-G	$4.0 \cdot 10^{-6}$	0.462	$5.94 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$	0.453	$2.80 \cdot 10^{-5}$	$4.4 \cdot 10^{-6}$	0.442	$1.35 \cdot 10^{-4}$	-	-	$4.79 \cdot 10^{-4}$
N3-G	$1.1 \cdot 10^{-4}$	0.283	$4.70 \cdot 10^{-6}$	$3.8 \cdot 10^{-5}$	0.279	$2.10 \cdot 10^{-5}$	$1.5 \cdot 10^{-4}$	0.279	$9.30 \cdot 10^{-5}$	-	-	$3.01 \cdot 10^{-4}$
C	0.913	0.057	>0.999	0.954	0.077	>0.999	0.975	0.095	>0.999	0.944	0.082	>0.999
Enol-C	0.087	0.130	$3.33 \cdot 10^{-18}$	0.046	0.179	$5.95 \cdot 10^{-16}$	0.025	0.216	$1.04 \cdot 10^{-13}$	0.056	0.195	$6.19 \cdot 10^{-12}$
N3-C	$1.7 \cdot 10^{-4}$	0.102	$1.88 \cdot 10^{-6}$	$1.5 \cdot 10^{-5}$	0.141	$1.11 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$	0.170	$6.39 \cdot 10^{-5}$	$2.0 \cdot 10^{-5}$	0.156	$2.52 \cdot 10^{-4}$
T	0.968	1.727	>0.999	0.990	1.721	>0.99	0.996	1.798	>0.999	>0.999	1.738	>0.999
2-Enol-T	$8.7 \cdot 10^{-6}$	1.842	$1.29 \cdot 10^{-11}$	$9.0 \cdot 10^{-6}$	1.854	$2.95 \cdot 10^{-10}$	$8.8 \cdot 10^{-6}$	1.958	$6.61 \cdot 10^{-9}$	$8.7 \cdot 10^{-6}$	1.912	$7.77 \cdot 10^{-8}$
4-Enol-T	0.032	3.252	$1.23 \cdot 10^{-8}$	0.010	3.274	$1.19 \cdot 10^{-7}$	0.004	3.462	$1.16 \cdot 10^{-6}$	$1.8 \cdot 10^{-4}$	3.383	$7.07 \cdot 10^{-6}$

The nucleosides are tautomer stable; all three different analyses - the energetics, the correlation between calculated and experimental chemical shifts, and the NMR population fit - are in agreement that the Watson-Crick tautomers are the dominant species. The NMR and energetics are not in agreement for the temperature dependent trends: the energetics suggest an increase of the minor tautomer fractions with temperature, the NMR a decline of their fractions. Considering the small absolute values of the minor tautomer fractions and the estimated errors of the fit, this effect may be an artifact and should not distract from the good agreement between both methods. Overall, from the results of all three methods combined, it can be expected that the nucleosides only present the Watson-Crick tautomer hydrogen bonding pattern at the investigated temperatures in solution. This is also consistent with the results for the plain nucleobases presented in chapter 4.5. In the literature, an energy difference between the N1- and N3-guanosine-tautomers of 2.9 kcal/mol was reported, and an difference of 1.4 kcal/mol to the enol-tautomer,[326] which is lower than the EC-RISM result for 298.15 K of 6.38 (N3), respectively 6.21 (enol) kcal/mol. For adenosine, minor tautomer fractions of less than 10^{-5} are presented,[327] and for cytidine no minor tautomer populations could be measured,[328] this is also in agreement with EC-RISM. The tautomeric situation for the natural nucleobases and the nucleosides seems to be clarified during this work. The next step is the addition of a phosphate chain and the investigation of the nucleoside mono- and triphosphates.

4.7.3 Nucleotides

The addition of a phosphate chain and a charge (due to the fact that all phosphate chains are fully deprotonated in this work) substantially affects the tautomerism of the nucleoside monophosphates (Table 43) compared to the plain nucleobases or nucleosides. Therefore, the single species have to be discussed in detail.

Table 43: Results for the nucleoside monophosphates at 278.15 and 363.15 K. The investigated tautomers are given in the first column; the reaction free energies $\Delta_r G$ (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM and MP2/6-311+G(d,p)/EC-RISM^T levels of theory are shown in columns 2-5. The corresponding populations are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. For GMP and its tautomers the values at 368 K instead of 363 K are shown due to convergence issues at 363 K.

ΔG / population	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	$\Delta_r G$	x_{taut}	x_{taut}	x_{taut}	x_{taut}
	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T
	278.15 K	278.15 K	363.15 K	363.15 K	278.15 K	278.15 K	363.15 K	363.15 K
AMP	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
N1-AMP	3.37	3.14	5.62	6.13	0.002	0.003	$4.1\cdot10^{-4}$	$2.0\cdot10^{-4}$
N3-AMP	3.23	3.05	4.37	4.77	0.003	0.004	0.002	0.001
GMP	0.00	0.00	0.00	0.00	0.180	0.169	0.482	0.509
Enol-GMP	-0.84	-0.88	-0.05	0.03	0.820	0.831	0.518	0.491
N3-GMP	17.46	17.43	12.40	12.46	$3.4\cdot10^{-15}$	$3.4\cdot10^{-15}$	$2.1\cdot10^{-8}$	$2.0\cdot10^{-8}$
CMP	0.00	0.00	0.00	0.00	0.962	0.951	0.987	>0.99
Enol-CMP	1.79	1.64	3.14	3.47	0.038	0.049	0.013	0.008
N3-CMP	7.04	7.03	5.35	5.36	$2.8\cdot10^{-6}$	$2.8\cdot10^{-6}$	$6.0\cdot10^{-4}$	$5.9\cdot10^{-4}$
TMP	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
2-Enol-TMP	12.36	12.33	9.87	9.93	$1.9\cdot10^{-10}$	$2.1\cdot10^{-10}$	$1.1\cdot10^{-6}$	$1.1\cdot10^{-6}$
4-Enol-TMP	5.31	5.31	4.00	3.99	$6.7\cdot10^{-5}$	$6.7\cdot10^{-5}$	0.004	0.004
UMP	0.00	0.00	0.00	0.00	>0.99	>0.99	>0.99	>0.99
2-Enol-UMP	14.48	14.45	11.55	11.61	$4.2\cdot10^{-12}$	$4.4\cdot10^{-12}$	$1.1\cdot10^{-7}$	$1.0\cdot10^{-7}$
4-Enol-UMP	11.58	11.58	9.20	9.20	$8.0\cdot10^{-10}$	$8.0\cdot10^{-10}$	$2.9\cdot10^{-6}$	$2.9\cdot10^{-6}$

The AMP is slightly less tautomer-stable than the adenine or adenosine, and the ranking of the minor tautomers is switched. This switching of the minor tautomers is also observed in the NMR fit of adenosine (Table 42). At 278 K, the reaction free energies relative to the Watson-Crick tautomer are around 3 kcal/mol regardless whether EC-RISM or EC-RISM^T is used, leading to minor tautomer fractions of approximately 0.1. The picture changes at 363 K, where the reaction free energies increase significantly, especially when looking at the EC-RISM^T results: here, the Watson-Crick tautomer is stable with very small minor tautomer fractions. This is in contrast to the expectations (from Boltzmann statistics, where larger minor tautomer fractions are obtained for the same reaction free energy at higher temperatures) that the minor tautomers should be stabilized at higher temperatures. Further investigation of the main conformations of the AMP tautomers (Figure 36) showed that the configuration of the C1' switched for the Watson-Crick tautomer compared to the input structure. The main conformation of the Watson-Crick tautomer is not one of the two conformations considered in the investigation of AMP at high-pressures (chapter 4.6), but the goal in this chapter was to determine which of both of the conformations is the main species, no sampling was done there. The results of this high-pressure calculations may be improved by considering the full conformational ensemble calculated for the temperature dependent investigations. Nevertheless, the main conformation presented here is not well suited for the incorporation in a nucleic acid strand due to geometry constraints, but seems to be stable in free solution.

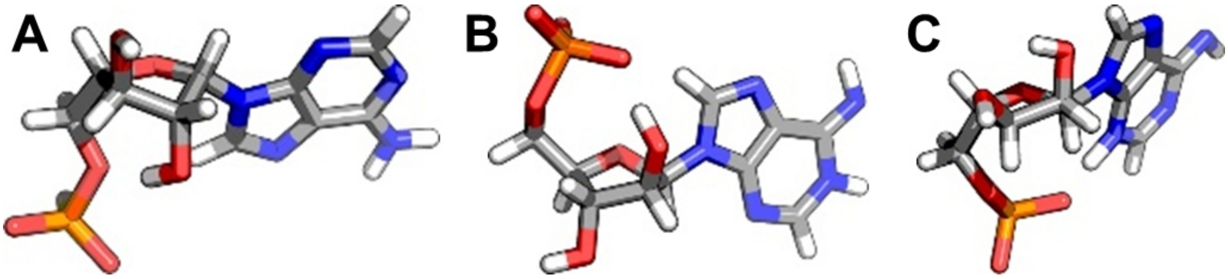


Figure 36: Main conformations of the AMP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, nearly 100%), N1-tautomer (B, 50-63%) and N3-tautomer (C, 54-58%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (6117, 3423 and 6456, respectively) are given in the structures subfolder of SI part 07.

Before discussing the GMP results, some convergence issues that occurred have to be reviewed. For three conformations of the enol tautomer, each at a specific temperature (with the numbers 8099 (at 293.15 K), 6239 (at 303.15 K), 1753 (at 363.15 K), the full results for each conformation at each temperature can be found in the Energies subfolder of SI part 07), the single conformation dominated the whole nucleotide energetically. This is because these EC-RISM calculations converged by chance: The convergence criterion of an EC-RISM calculation is an energy residuum of 0.01 kcal/mol, this was achieved by these calculations randomly. For example, the difference in intramolecular energy between iterations 2 and 3 for the conformation 8099 at 293.15 K is approximately 14 kcal/mol, and the difference in excess chemical potentials between these iterations is approximately -14 kcal/mol, resulting in a very low energy residuum matching the convergence criterion, while the calculation is not converged in a physical sense. Therefore, the data at

363 K is not presented here and replaced by the 368 K results, since all trends are monotonous with exception of the data with these three "wrong" calculations. At 278 K, the main tautomer of GMP is the enol with a fraction of closely 0.8; the N3 tautomer is not populated. The main conformation of the Watson-Crick tautomer has a favorable interaction between the H8, the H3' and the phosphate chain, while the main conformation of the enol tautomer has a totally different shape; here, the sugar conformation is flipped, leading to a more ring-like structure between the sugar and the N3. In the main N3-conformation, the base points towards the sugar, a flip from the *anti*-conformation in the input structure to a *syn*-conformation happened here. By increasing the temperature, the fraction of the N1 tautomer increases steadily while the enol fraction is similarly declining. The N3 tautomer is still not populated at 368 K, while the enol fraction declines to 0.49 using EC-RISM^T. For the GMP, the Watson-Crick tautomer is stabilized by pressure, but it is not the main tautomer at ambient conditions since the enol-tautomer is stabilized by hydrogen-bonds. In the literature, the N1-tautomer is reported to be the main species; it is favored to the N3-tautomer by 17.5 kcal/mol and to the enol-tautomer by 9.4 kcal/mol.[326] But the study was done for the monoanion, in contrast the dianion is considered here. This shows that for a complete view of the tautomeric behavior of nucleotides, the consideration of all ionization states is needed. The computational effort needed for this is large but important for future projects.

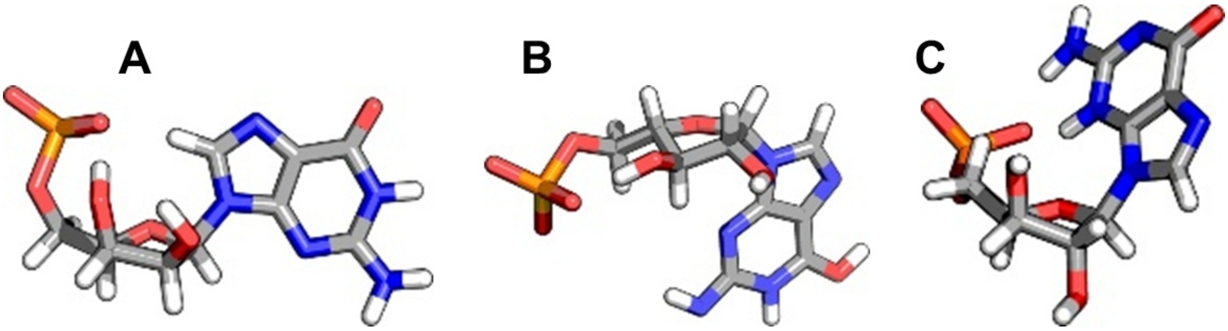


Figure 37: Main conformations of the GMP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 100-98%), enol-tautomer (B, 73-52%) and N3-tautomer (C, 67-45%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (3111, 4182 and 4134, respectively) are given in the structures subfolder of SI part 07.

The N1-GMP main conformation reported for the monoanion in the literature (Ref. 326) is more folded than the one presented here for the dianion, but this is expected since in the literature the PCM solvation model is used which often favors more folded conformations due to the strong effect of the cavitiy term.

The energetic differences between CMP and enol-CMP are very small with only approximately 1.7 kcal/mol at 278.15 K, while the N3 tautomer plays a minor role. This results in fractions of 0.04 (EC-RISM) and 0.05 (EC-RISM^T), respectively. Like in AMP, the Watson-Crick tautomer is stabilized at higher temperatures, leading to per thousand fractions of the minor tautomers at 363.15 K and reaction free energies of approximately 3.3 (enol-CMP), respectively 5.3 kcal/mol (N3-CMP). In the literature, no evidence of a minor CMP tautomer was found via UV-Raman-spectroscopy.[329] The main conformation of the Watson-Crick tautomer is stretched, while the one of the enol is folded. The N1-tautomer main conformation has, like the AMP Watson-Crick tautomer main species, a flipped C1' configuration.

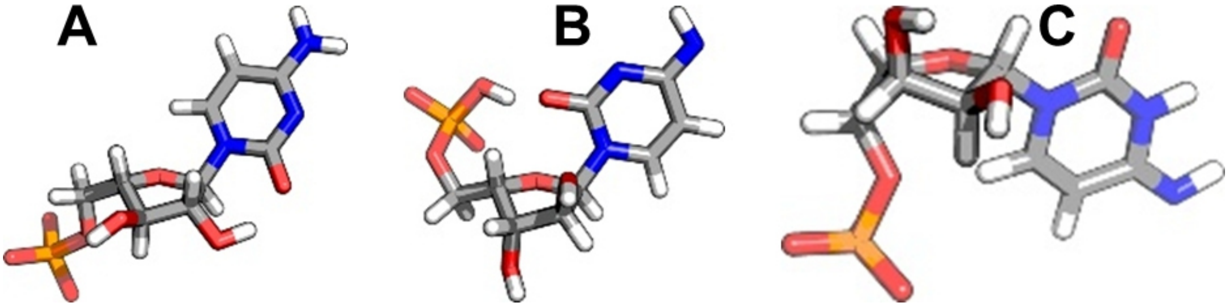


Figure 38: Main conformations of the CMP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 75-60%), enol-tautomer (B, nearly 100%) and N1-tautomer (C, 72-63%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (6038, 914 and 5016, respectively) are given in the structures subfolder of SI part 07.

The TMP is tautomer stable regardless of the temperature with high reaction free energies of the minor tautomers (approximately 12.3 (enol2-TMP) and 5.3 kcal/mol (enol4-TMP)). The energetic differences are getting smaller (ca. 10 and 4 kcal/mol enol2/enol4-TMP) with higher temperatures, but the Watson-Crick tautomer remains the dominant one. The main conformations of the Watson-Crick- and 2-enol-tautomer are mainly different w.r.t. the sugar configuration and the orientation of the phosphate chain, while the 4-enol-tautomer shows the switch in C1' configuration.

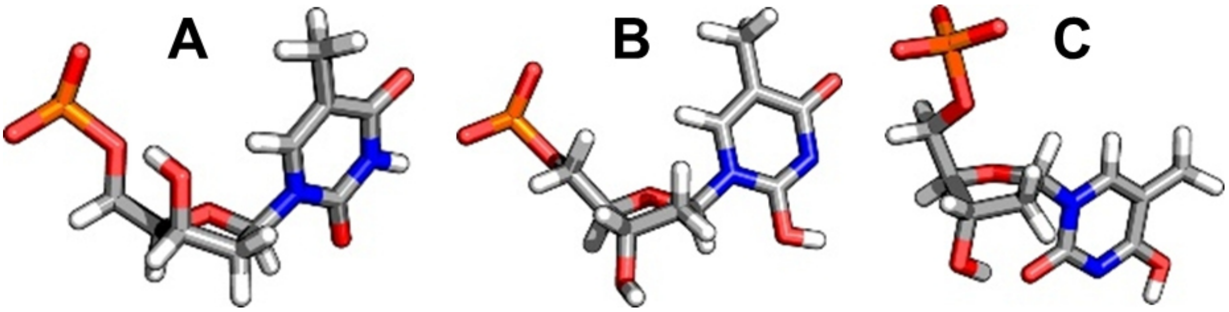


Figure 39: Main conformations of the TMP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 69-70%), 2-enol-tautomer (B, 98-93%) and 4-enol-tautomer (C, 98-97%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (52, 104 and 3933, respectively) are given in the structures subfolder of SI part 07.

The UMP shows the same behavior, but with a stronger favorization of the Watson-Crick tautomer, which makes it special within the nucleoside monophosphates, since it is the only species which is tautomer stable over the whole temperature range. The UMP reaction free energies are around 14-11.5 kcal/mol (278-363 K) for the enol2- and 11.5-9.2 kcal/mol for the enol4-tautomer. All of the main conformations show the switch in C1' configuration. The stability of the RNA base is higher than the one of the corresponding DNA base, especially at high temperatures. This can be a hint that the RNA may be the first of both nucleic acids developed at hydrothermal vents.

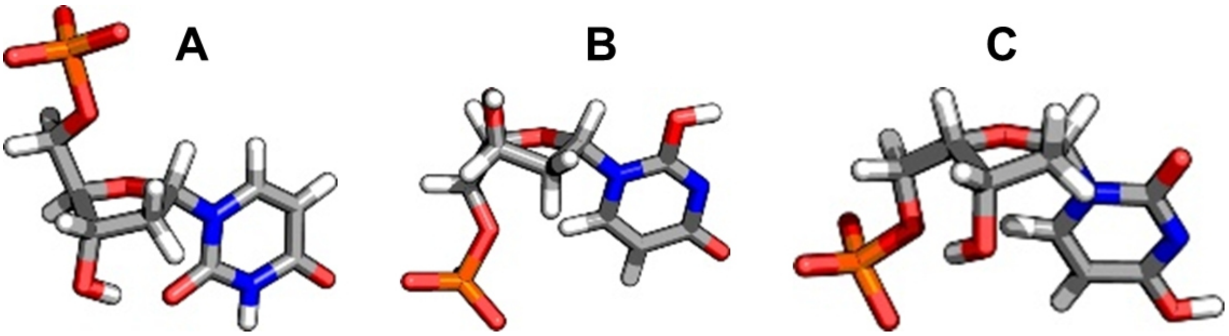


Figure 40: Main conformations of the UMP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 64-55%), 2-enol-tautomer (B, 99-96%) and 4-enol-tautomer (C, 51%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (341, 502 and 113, respectively) are given in the structures subfolder of SI part 07.

The main difference between the tautomer stable nucleobases and nucleosides and the more unstable nucleoside monophosphates is the option of the monophosphates to form conformations with strong intramolecular hydrogen bonds which can stabilize rare tautomers, as can be seen especially for the GMP.

But these kind of intramolecular hydrogen bonds are most likely not possible in nucleic acid strands, since they form a chain with their phosphate groups.

NMR experiments were only performed for the nucleosides and nucleoside triphosphates, not for the monophosphates, but NMR calculations of the triphosphates were not possible due to the high computational effort (only energy calculations were possible for these species) and especially the storage demand needed for these calculations. To get maximum benefit from the calculations and experiments, it was therefore decided to compare the experimental chemical shifts of the triphosphates with the calculated ones for the monophosphates. This way the experimental data can be used together with calculations in a similar fashion as it was done for the histamine and the unphosphorylated nucleosides. This approximation presupposes a similar behavior of the mono- and triphosphates not only from the tautomer perspective, but also from the conformational behavior and the influence of the phosphate chain, the charge and the environmental conditions on the chemical shifts. Thus, a comparison of the energetics of the nucleoside mono- and triphosphates has to be done.

Table 44: Results for the nucleoside triphosphates at 278.15 and 363.15 K. The investigated tautomers are given in the first column; the reaction free energies $\Delta_r G$ (in kcal/mol), at the MP2/6-311+G(d,p)/EC-RISM and MP2/6-311+G(d,p)/EC-RISM^T levels of theory are shown in columns 2-5. The corresponding populations are given in columns 6-9. The data comprises Boltzmann averaging of the energies of all rotamers for a specific tautomer. For UTP and its tautomers, the values at 368 instead of 363 K are presented.

ΔG /	$\Delta_\mathrm{r}G$	$\Delta_\mathrm{r}G$	$\Delta_\mathrm{r}G$	$\Delta_\mathrm{r}G$	x_taut	x_taut	x_taut	x_taut
population	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T	EC-RISM	EC-RISM^T
	278.15 K	278.15 K	363.15 K	363.15 K	278.15 K	278.15 K	363.15 K	363.15 K
ATP	0.00	0.00	0.00	0.00	$1.5\cdot10^{-5}$	$1.5\cdot10^{-5}$	0.002	0.002
N1-ATP	-6.15	-6.16	-4.45	-4.43	>0.99	>0.99	>0.99	>0.99
N3-ATP	1.30	1.32	0.31	0.26	$1.4\cdot10^{-6}$	$1.3\cdot10^{-6}$	0.001	0.002
GTP	0.00	0.00	0.00	0.00	0.789	0.801	0.445	0.413
Enol-GTP	0.73	0.77	-0.16	-0.25	0.210	0.198	0.552	0.584
N3-GTP	4.33	4.32	3.72	3.71	$3.1\cdot10^{-4}$	$3.2\cdot10^{-4}$	0.003	0.002
CTP	0.00	0.00	0.00	0.00	$1.0\cdot10^{-6}$	$1.1\cdot10^{-6}$	$2.1\cdot10^{-4}$	$1.8\cdot10^{-4}$
Enol-CTP	-0.79	-0.79	0.23	0.21	$4.2\cdot10^{-6}$	$4.5\cdot10^{-6}$	$1.5\cdot10^{-4}$	$1.4\cdot10^{-4}$
N3-CTP	-7.63	-7.60	-6.12	-6.21	>0.99	>0.99	>0.99	>0.99
dTTP	0.00	0.00	0.00	0.00	0.978	0.979	0.819	0.807
2-Enol-dTTP	2.10	2.13	1.09	1.03	0.022	0.021	0.181	0.193
4-Enol-dTTP	35.72	35.72	26.23	26.25	$8.5\cdot10^{-29}$	$8.5\cdot10^{-29}$	$1.3\cdot10^{-16}$	$1.3\cdot10^{-16}$
UTP	0.00	0.00	0.00	0.00	0.015	0.015	0.108	0.105
2-Enol-UTP	2.07	2.10	2.00	2.03	$3.6\cdot10^{-4}$	$3.4\cdot10^{-4}$	0.039	0.043
4-Enol-UTP	-2.30	-2.30	-2.24	-2.23	0.984	0.984	0.848	0.847

The results of the calculations of the nucleoside triphosphates are given in Table 44. There are major differences between these results and the ones for the nucleoside monophosphates. In contrast to AMP, the main tautomer of the ATP is not the Watson-Crick- but instead the N1-tautomer. The main conformations of both tautomers have a hydrogen bond between H8 of the base and O5', while the one of the N3-tautomer is more folded. The minor tautomers have increasing populations with temperature, but the N1-tautomer is dominant even at 363 K with a reaction free energy of 4.43, respectively 4.69 kcal/mol to the Watson-Crick- and N3-tautomer.

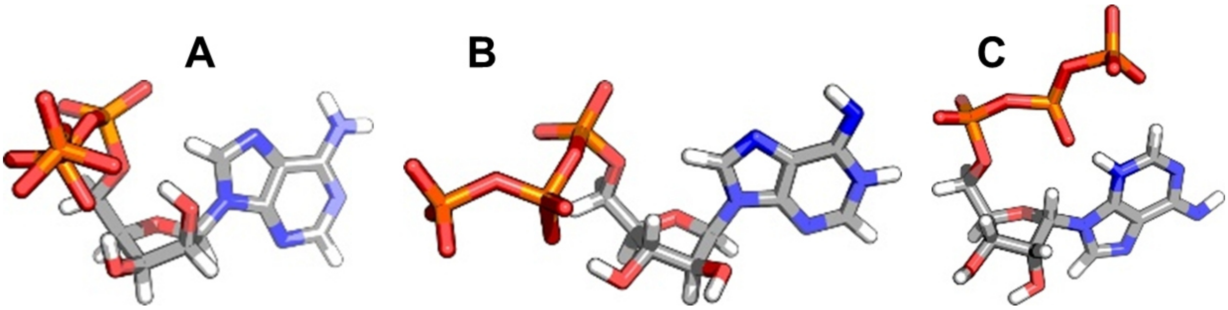


Figure 41: Main conformations of the ATP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 33-54%), N1-tautomer (B, nearly 100%) and N3-tautomer (C, 98-95%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (329, 215 and 6473, respectively) are given in the structures subfolder of SI part 07.

For GMP, the enol-tautomer is the main species at low temperatures; at high temperatures, it is nearly equally weighted with the Watson-Crick tautomer. This is a disparity to the GTP, where the N1-tautomer is dominant, with an increasing fraction of the minor enol-tautomer at rising temperatures. The reaction free energies to the N3-tautomer are relatively small compared to the guanine. The main conformation of the Watson-Crick-tautomer has favorable hydrogen bonding interactions between H8 and the phosphate chain and this chain and the 3'OH group; in contrast to the enol- and N3-tautomers, the base is in an *anti*-configuration. The enol- and N3-tautomer main conformations show more interactions between the base and the sugar/phosphate chain, due to a *syn*-configuration of the base.

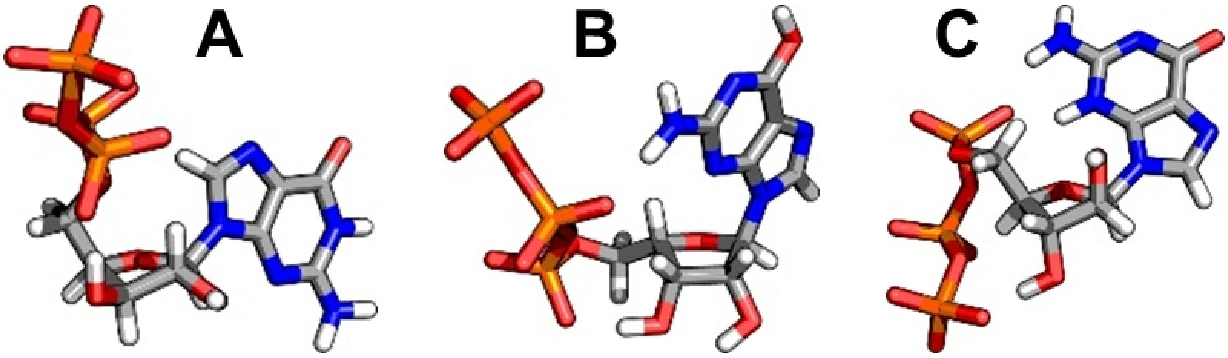


Figure 42: Main conformations of the GTP tautomers as obtained from EC-RISM^{T} calculations: the Watson-Crick tautomer (A, 64-69%), enol-tautomer (B, nearly 100%) and N3-tautomer (C, nearly 100%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (2583, 4182 and 4134, respectively) are given in the structures subfolder of SI part 07.

While the N3-tautomer is the least populated form of the monophosphate, it is the dominant species of the CTP. The minor tautomers, the Watson-Crick- and enol-tautomer, have an expanding fraction with rising temperature and a switch in the ranking: The Watson-Crick tautomer is the least abundant form at low temperatures whereas it becomes the second most abundant form at high temperatures. The main conformations of the tautomers are important since they differ strongly from the input structures of the conformational sampling workflow. The input structures have the "standard" stereochemistry: an α-D-ribose-configuration with a 4*S*-configuration is used. The main conformations of the enol- and N1-tautomers have the sugar switched from the 4*S*- to the 4*R*-configuration, and the main conformation of the Watson-Crick tautomer from the α-D-ribose- to the β-D-ribose-configuration.

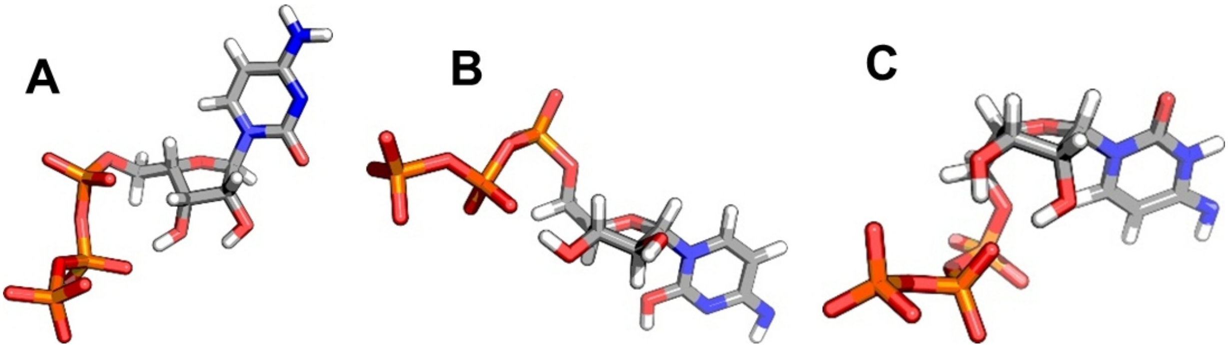


Figure 43: Main conformations of the CTP tautomers as obtained from EC-RISM^{T} calculations: the Watson-Crick tautomer (A, 90-87%), enol-tautomer (B, 99-97%) and N1-tautomer (C, 73-86%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (4751, 914 and 5016, respectively) are given in the structures subfolder of SI part 07.

While the dTTP main tautomer is the Watson-Crick-tautomer, the 2-enol-fraction is rising with temperature, from 2% at 278.15 K to 19% at 363.15 K. Similarly, the TMP Watson-Crick tautomer is the dominant species of the monophosphate, but followed by a minor fraction of the 4-enol-tautomer, which is strongly disfavored for the triphosphate. From the main conformations, the Watson-Crick- and 4-enol-tautomer both have a ring-like structure due to hydrogen bonding between the sugar and phosphate-chain; the first with, the latter without a switch in the C1' configuration. The 2-enol-tautomer main conformation forms a ring between the base and the phosphate-chain, which seems to be extremely stable at high temperatures.

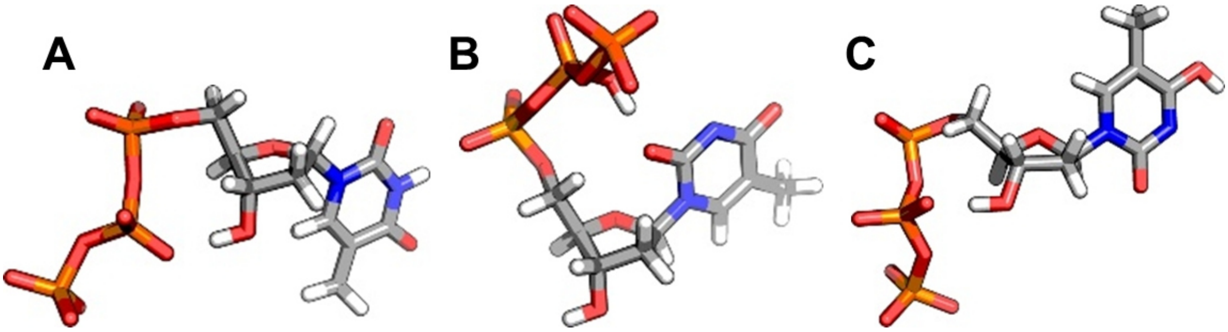


Figure 44: Main conformations of the dTTP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 96-95%), 2-enol-tautomer (B, nearly 100%) and 4-enol-tautomer (C, 85-68%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (540, 1539 and 1329, respectively) are given in the structures subfolder of SI part 07.

The most stable nucleoside monophosphate is the UMP, with reaction free energies in the range of 10 kcal/mol at high temperatures. The UTP is less stable, with the largest minor tautomer fraction being 0.015 at 278 K (Watson-Crick tautomer) increasing to 0.105 at 368 K, but the main tautomer is, in contrast to UMP, the 4-enol-tautomer. For UTP, the same convergence issues as for GMP occurred at 328, 338, 343 and 363 K, leading to a fully populated enol-4-tautomer, so as for the GMP, the values at 368 K are presented here. The main conformation of the Watson-Crick tautomer has a ring between the sugar and the phosphate-chain, as well as the inversion of the C1' configuration. The 2-enol- and 4-enol-tautomer main species form a ring between sugar, base and phosphate-chain. The 2-enol-tautomer main conformation is very similar to that of the respective dTTP tautomer.

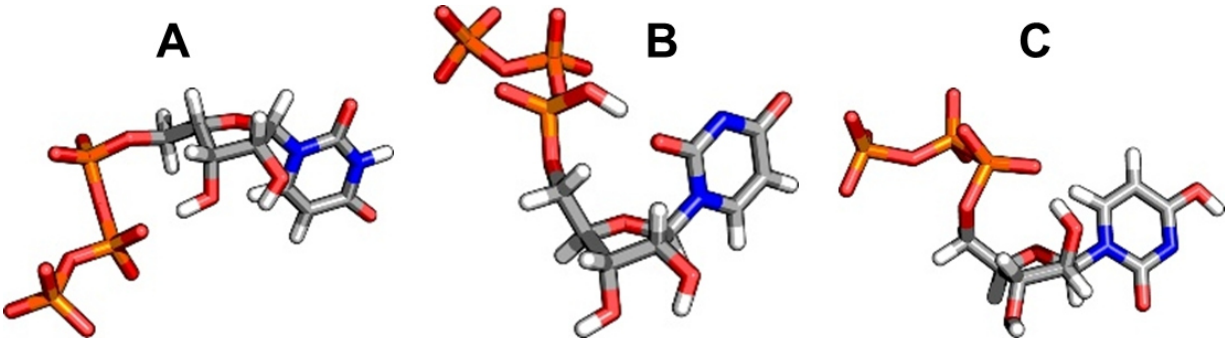


Figure 45: Main conformations of the UTP tautomers as obtained from EC-RISM^T calculations: the Watson-Crick tautomer (A, 74-66%), 2-enol-tautomer (B, nearly 100%) and 4-enol-tautomer (C, 91-83%). The fractions are the conformational fractions within the respective tautomer at 278.15 and 372.756 K (the trends are monotonous with temperature changes). The structures (2494, 638 and 1077, respectively) are given in the structures subfolder of SI part 07.

All nucleoside triphosphates behave energetically different from the monophosphates. But while being not fully comparable from this point of view, the calculated chemical shifts of the nucleoside monophosphate bases, which are only indirect, via the calculated conformer populations, dependent on the energetics of the tautomers, can still be used to extract the tautomer ratios of the triphosphates by fitting these calculated chemical shifts to the experimental shifts of the triphosphates. This requires the assumption that the two additional phosphates in the phosphate chain are not affecting the conformational ensemble and the chemical shifts of the base atoms strongly, which would be possible for example by the formation of additional hydrogen bonds formed with the triphosphate chain or if the CSA of the additional phosphates influences the chemical shifts strongly. This is investigated in the following section.

The correlation between the experimental shifts, measured at different temperatures (their assignment to the respective atoms of the nucleotides is shown in Table 36 to Table 39) of the nucleoside triphosphates and the calculated shifts of the monophosphates is shown in Table 45. The correlation between experiment and calculation is, compared to the results for the nucleosides, low. The absolute numbers of the shifts are so high that a graphic presentation of data (similar to Figure 35) is not presented here (the full data is given in the NMR subfolder of SI part 07). The error metrics are very high, but worth discussing. All metrics, RMSE, MAE, MSE, slope, intercept and coefficient of determination, indicate a dominance of the Watson-Crick tautomer for AMP/ATP, which is in agreement with the energy calculations for AMP, but in contrast to those for ATP. Since the ATP is an outlier in the series of adenosine, AMP and ATP, this result is consistent with all data except the ATP results. For GMP/GTP, the error metrics and linear regression parameters suggest the N3-tautomer to be the main species followed by the Watson-Crick tautomer. All the metrics are worse for the enol-tautomer. This is opposed to the energetics, which result in the Watson-Crick- and enol-tautomer being the main forms of the GMP, and the Watson-Crick-tautomer as the favored species of GTP. But since the errors are high, and the difference between the metrics for the Watson-Crick- and N3-tautomers is, compared to the AMP/ATP tautomers, small, these evidence for the preferred tautomers are rather weak. The situation for CMP/CTP is similar to the one for AMP/ATP: all metrics signal a dominance of the Watson-Crick tautomer. This is in agreement with the CMP energetics but contrary to the energetics of CTP. For UMP/UTP, the metrics indicate the dominance of the 2-enol-tautomer, while the energetics result in the Watson-Crick-tautomer being the main tautomer of UPM and the 4-enol-tautomer the one of the triphosphate.

Table 45: Error metrics (RMSE, MSE and MAE in ppm) of the nucleoside monophosphate base ^{1}H, ^{13}C and ^{15}N NMR chemical shifts calculated for the tautomers to the respective corresponding experimental data of the triphosphates (calculations and experiments performed at 278, 298, 323 and 348 K) and the linear regression data slope m, intercept b in ppm and coefficient of determination (R^2) are shown.

Error	RMSE	MAE	MSE	*m*	*b*	R^2
	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K	(278 K/298 K
	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)	323 K/348 K)
AMP/ATP	17.36/18.16	13.99/14.65	-13.99/-14.65	0.993/0.990	14.97/16.09	0.980/0.978
	19.94/21.33	16.17/17.21	-16.17/-17.21	0.981/0.977	18.78/20.38	0.974/0.970
N1-AMP/ATP	141.4/141.3	72.99/72.96	-71.59/71.44	0.925/0.922	81.98/82.35	0.229/0.229
	141.8/141.3	73.09/72.89	-71.44/-71.13	0.903/0.912	84.99/83.44	0.222/0.227
N3-AMP/ATP	165.6/165.7	80.02/79.99	-78.79/-78.73	0.818/0.815	104.1/104.4	0.141/0.140
	166.4/166.0	80.14/79.98	-78.85/-78.68	0.794/0.806	107.5/105.7	0.134/0.139
GMP/GTP	90.57/90.49	59.39/59.31	-58.22/-58.13	1.283/1.283	18.84/18.74	0.527/0.528
	90.34/90.14	59.18/58.99	-57.98/-57.77	1.283/1.283	18.54/18.39	0.530/0.532
Enol-GMP/GTP	150.2/150.3	85.16/85.11	-85.16/-85.11	0.561/0.559	146.3/146.5	0.062/0.062
	150.3/150.4	85.01/84.85	-85.01/-84.85	0.558/0.557	146.6/146.6	0.062/0.062
N3-GMP/GTP	65.13/65.01	49.86/49.73	-49.86/-49.73	1.252/1.252	14.77/14.66	0.756/0.757
	64.81/64.55	49.52/49.26	-49.52/-49.26	1.251/1.250	14.50/14.43	0.759/0.760
CMP/CTP	25.80/28.83	22.90/25.92	-22.90/-25.92	1.042/1.035	18.17/21.94	0.972/0.968
	35.97/49.79	32.98/46.33	-32.98/-46.33	1.023/1.004	30.37/45.88	0.957/0.930
Enol-CMP/CTP	195.8/195.0	111.0/110.6	-111.0/-110.6	1.270/1.286	80.30/78.02	0.212/0.219
	194.5/192.9	110.4/110.4	-110.4/-110.4	1.286/1.288	77.89/77.56	0.221/0.224
N3-CMP/CTP	44.64/51.00	38.20/44.81	-32.51/-39.50	0.720/0.671	64.32/76.95	0.787/0.772
	60.90/76.32	54.14/67.85	-49.51/-64.47	0.603/0.508	94.72/120.6	0.734/0.649
UMP/UTP	94.06/146.2	88.38/137.3	-88.38/-137.3	0.927/0.823	96.51/157.0	0.770/0.530
	343.9/2010	322.2/1877	-322.2/-1877	0.440/-2.769	384.6/2296	0.055/0.062
2-Enol-UMP/UTP	52.41/60.35	49.23/57.23	-49.23/-57.23	0.978/0.929	51.74/65.18	0.922/0.908
	80.37/107.3	76.40/101.1	-76.40/-101.1	0.810/0.657	97.59/139.5	0.810/0.657
4-Enol-UMP/UTP	77.53/88.20	67.53/76.93	-67.53/-76.93	0.977/0.931	70.11/84.57	0.723/0.651
	114.7/149.8	99.58/128.9	-99.58/-128.9	0.820/0.676	119.6/164.9	0.461/0.251

The errors between the experimental shifts of the nucleoside-triphosphates and the calculated ones of the nucleoside-monophosphates are high. The correlation between them, measured with these error metrics and the linear regression parameters, is low. The closest agreement is between the Watson-Crick tautomers of AMP/ATP with an RMSE of 17.36 ppm, a slope of 0.993, an intercept of 14.97 ppm, and a coefficient of determination of 0.98 at 278.15K. Examples for a poor correlation are the enol-tautomer of CMP/CTP and the Watson-Crick tautomer of UMP/UTP at high temperatures. But besides these worse correlations, the results indicate that the Watson-Crick tautomers seem to be the most important species for the most nucleotides, which is in agreement with the results for the nucleosides and nucleobases. An explanation for this correlation, which is in contrast to the observations made for the nucleosides, is not only the different systems used for the experiment and calculations. Also the chemical shift anisotropy seems to play an important role since the calculations on the monophosphates are strongly influenced by the anisotropy of the phosphates.[6]

To investigate the influence of the phosphate CSA, in Table 46, the isotropic and anisotropic shielding constants of the main conformations of the N3-tautomer of adenosine and AMP at 298.15 K are shown. The isotropic shielding constant, as obtained from QC calculations, is the trace of the shielding tensor, while the CSA is computed from the shielding tensor as described in Ref. 330. This anisotropy should cancel in solution, and therefore in the experiment, but in the calculation, the absolute values of the CSA seem to influence the isotropic shielding constant. This phenomenon is investigated here on an example and needs to be extensively studied in the future, since it seems to be important for further NMR calculations. To allow the comparison of the calculated shielding constants of the single conformations with the corresponding experimental data, the shielding constants from the experiment are needed. Since they are not obtained by the experiment, they need to be calculated. This can be done in a reverse way of the calculation of computational chemical shifts, taking the experimental chemical shifts and subtract them from the calculated reference shielding of the respective nuclei, which are given in chapter 4.4. This results in the here called “experimental” shielding constants.

Table 46: Calculated isotropic (i) and anisotropic (a) shielding constants of the main conformations of the N3-tautomers of adenosine and AMP and the experimental shielding constants (e) for the systems, all at 298.15K. Also, the differences between the calculated isotropic and the experimental shielding constants (which is equivalent to the differences in chemical shifts) and the difference in anisotropic shielding constants between the N3-AMP and N3-adenosine main conformations are shown. The anisotropic shielding constants are obtained from QC calculations together with the isotropic ones, the theory needed for their calculation is described and discussed in Ref. 331.

Atom	$\sigma_{\text{N3-A(e)}}$	$\sigma_{\text{N3-A(i)}}$	$\sigma_{\text{N3-A(a)}}$	$\Delta\sigma_{\text{N3-A(e-i)}}$	$\sigma_{\text{N3-AMP(e)}}$	$\sigma_{\text{N3-AMP(i)}}$	$\sigma_{\text{N3-AMP(a)}}$	$\Delta\sigma_{\text{N3-AMP(e-i)}}$	$\Delta\sigma_{\text{(N3-AMP(a)-N3-A(a))}}$
H2	23.63	23.90	6.78	-0.27	23.39	22.11	8.32	1.28	1.54
H8	23.73	24.05	5.85	-0.33	23.66	21.16	10.48	2.51	4.63
C2	45.86	54.08	119.25	-8.22	45.37	53.11	122.73	-7.74	3.48
C4	49.90	63.53	110.38	-13.63	49.07	42.11	127.02	6.96	16.64
C5	79.22	71.03	95.28	8.20	79.59	11.24	181.02	68.35	85.75
C6	42.61	32.24	133.32	10.38	42.50	30.36	157.88	12.14	24.56
C8	57.89	60.53	79.85	-2.64	58.32	23.42	133.96	34.90	54.11
C1’	110.70	102.91	55.44	7.79	111.64	106.77	47.16	4.87	-8.28
C2‘	159.18	152.19	31.46	6.98	123.93	121.49	27.98	2.45	-3.48
C3‘	126.84	118.21	33.47	8.63	127.84	121.04	47.32	6.80	13.85
C4‘	113.53	104.47	48.45	9.06	114.16	108.42	55.72	5.74	7.27
C5‘	136.43	129.83	44.12	6.60	133.11	127.77	52.46	5.34	8.34
N1	52.84	31.79	276.06	21.05	51.93	-77.31	459.81	129.24	183.75
N3	61.13	124.98	144.09	-63.85	61.38	42.28	290.96	19.10	146.87
N6	199.84	107.09	184.15	92.75	200.12	-344.06	878.28	544.18	694.13
N7	48.21	47.71	307.57	0.50	46.03	-53.39	441.69	99.42	134.11
N9	105.10	83.25	103.59	21.85	109.28	72.76	112.14	36.52	8.55

The differences in the CSA are large between the calculations for adenosine and AMP. For most of the atoms, the CSA is larger for AMP than for adenosine (in some cases multiples of the adenosine CSA). This larger CSA results for the most atoms in a much smaller isotropic shielding constant (the exceptions are the C3’ and C4’ of the sugar, which can also be influenced by the difference in the sugar conformations introduced through the use of desoxyribonucleosides and ribonucleotides), and especially for some of the carbon and the nitrogen atoms in values out of the plausible range. This clearly shows that calculations on phosphorylated nucleosides with the EC-RISM/NMR workflow presented in this work are not reliable due to the large CSA effect introduced in the systems by the phosphate chain. The strong CSA effect of the phosphate chain was also an issue for the investigation of adenosine and the corresponding phosphorylated nucleosides at high-pressure (chapter 4.6). This effect needs to be studied in more detail in future work.

There are additional lessons learned from these results: the comparison of the calculated and experimental nitrogen shielding constants hints that the N3-tautomer is a minor species. For adenosine, the N3 shielding is largely overestimated computationally, since in this tautomer, it is, in contrast to the experiment, protonated and the addition of a hydrogen atom should result in a higher shielding for the nitrogen bound to it. Consequently, the N6 shielding is underestimated in the calculation since the proton is missing here. Further, the calculated shielding constants are in good agreement with the experiment, especially for the protons with less than half a ppm. The larger deviations for some of the carbon and nitrogen shielding constants can be explained by the fact that only a single conformation is considered here. Comparing the experimental shielding constants of the adenosine and the AMP shows that these experimental values are only slightly influenced by the addition of the phosphate chain. All of the results except the C2', N7 and N9 shielding constants differ by maximum 1 ppm (the N7 and N9 are not involved in the tautomerization and close to the C2' for which the change can be explained by the addition of the hydroxyl-group since desoxyribonucleosides and ribonucleotides are investigated). Besides this similarity in the experiments, the differences between the calculated shielding constants for adenosine and AMP are large. Also, a high correlation between calculation and experiment is observed for the adenosine and a low one for the AMP, so the calculated shielding constants of the AMP seem to be problematic.

The calculated shielding constant of the N3-AMP main conformation N3 fits the experimental value quite well; this could imply that the N3 protonation is correct, but since the CSA reduces all of the nitrogen shielding constants, this looks correct by accident. Most of the other nitrogen signals are strongly differing from the experiment, and especially N1, N6 and N7 results are implausible, therefore, the NMR fit may not be reliable for the phosphorylated compounds in general.

Although experiment and calculation do not match well, the calculated chemical shifts of the monophosphates were used here for the extraction of the tautomer fractions of the triphosphates, even though the analysis of the influence of the CSA hints that the results are not reliable. The results of the fitting process, which was performed as described in chapter 3.2.3.2, are given in Table 47.

Table 47: Results of the calculation of nucleoside monophosphate tautomer fractions from the experimental NMR chemical shifts of the corresponding triphosphates (NMR fit) at different temperatures. The results were obtained by using the ^1H and ^{15}N chemical shifts of the base only. The chemical shifts were normalized applying Eqn. 197. Additionally, the EC-RISM^T results calculated for the respective mono- (MP) and triphosphates are given.

Population	x_{taut}(278 K) NMR fit	x_{taut}(278 K) EC-RISM MP/TP	x_{taut}(298 K) NMR fit	x_{taut}(298 K) EC-RISM MP/TP	x_{taut}(323 K) NMR fit	x_{taut}(323 K) EC-RISM MP/TP	x_{taut}(348 K) NMR fit	x_{taut}(348 K) EC-RISM MP/TP
AMP/ATP	0.749	0.993/ $1.5\cdot 10^{-5}$	0.739	0.995/ $6.9\cdot 10^{-5}$	0.727	0.997/ $3.3\cdot 10^{-4}$	0.708	0.998/ 0.001
N1-AMP/ATP	0.251	0.003/ >0.999	0.261	0.002/ >0.999	0.273	$7.4\cdot 10^{-4}$/ >0.999	0.292	$4.5\cdot 10^{-4}$/ 0.998
N3-AMP/ATP	$1.9\cdot 10^{-5}$	0.004/ $1.4\cdot 10^{-6}$	$2.7\cdot 10^{-5}$	0.004/ $1.2\cdot 10^{-5}$	$1.8\cdot 10^{-4}$	0.002/ $1.1\cdot 10^{-4}$	$6.4\cdot 10^{-6}$	0.002/ $4.6\cdot 10^{-4}$
GMP/GTP	$1.6\cdot 10^{-5}$	0.169/ 0.789	$1.4\cdot 10^{-5}$	0.248/ 0.711	$1.3\cdot 10^{-5}$	0.352/ 0.588	$1.2\cdot 10^{-5}$	0.445/ 0.473
Enol-GMP/GTP	$3.4\cdot 10^{-7}$	0.831/ 0.210	$3.4\cdot 10^{-7}$	0.752/ 0.288	$3.4\cdot 10^{-7}$	0.648/ 0.411	$3.4\cdot 10^{-7}$	0.555/ 0.525
N3-GMP/GTP	>0.999	$3.4\cdot 10^{-15}$/ $3.1\cdot 10^{-4}$	>0.999	$4.2\cdot 10^{-13/}$ $6.3\cdot 10^{-4}$	>0.999	$4.7\cdot 10^{-11}$/ 0.001	>0.999	$9.6\cdot 10^{-10}$/ 0.002
CMP/CTP	0.766	0.951/ $1.0\cdot 10^{-6}$	0.759	0.966/ $3.9\cdot 10^{-6}$	0.748	0.981/ $2.8\cdot 10^{-5}$	0.754	0.989/ $7.9\cdot 10^{-5}$
Enol-CMP/CTP	0.047	0.049/ $4.2\cdot 10^{-6}$	0.082	0.034/ $1.3\cdot 10^{-5}$	0.113	0.019/ $3.9\cdot 10^{-5}$	0.164	0.011/ $7.7\cdot 10^{-5}$
N3-CMP/CTP	0.187	$2.8\cdot 10^{-6}$/ >0.999	0.159	$1.6\cdot 10^{-5}$/ >0.999	0.139	$8.4\cdot 10^{-5}$/ >0.999	0.082	$3.0\cdot 10^{-4}$/ >0.999
UMP/UTP	$4.5\cdot 10^{-6}$	>0.999/ 0.015	0.199	>0.999/ 0.034	0.070	>0.999/ 0.055	$5.1\cdot 10^{-5}$	>0.999/ 0.083
2-Enol-UMP/UTP	0.831	$4.4\cdot 10^{-12}$/ $3.6\cdot 10^{-4}$	0.747	$1.0\cdot 10^{-10}$/ $3.2\cdot 10^{-4}$	0.839	$2.5\cdot 10^{-9}$/ 0.006	0.900	$1.9\cdot 10^{-8}$/ 0.019
4-Enol-UMP/UTP	0.169	$8.0\cdot 10^{-10}$/ 0.984	0.054	$9.9\cdot 10^{-9}$/ 0.964	0.091	$1.3\cdot 10^{-7}$/ 0.939	0.100	$7.0\cdot 10^{-7}$/ 0.898

The results for the AMP/ATP system are in agreement with the error and linear regression metrics shown in Table 45: the Watson-Crick tautomer is the main species with fractions of 0.749 (278.15 K) decreasing to 0.708 (348.15 K) upon heating. This is in contrast to the energy calculations for the ATP, where the N1-tautomer is the most abundant species, but agrees with the trends for adenosine and AMP, although with a lower Watson-Crick-tautomer fraction.

The GMP/GTP system seems to be dominated by the error metrics given in Table 45 since the N3-tautomer, which matches the experimental results best, is fully populated. This is in disagreement with all other calculations for guanosine, GMP and GTP.

For CMP/CTP, the Watson-Crick-tautomer is the main form but slightly decreasing with temperature. The minor tautomers, the enol- and N3-tautomer, show switching of their population upon the heating process, the N3-tautomer fraction is declining from 0.187 at low to 0.082 at high temperatures, while the enol-fraction rises from 0.047 (278.15 K) to 0.164 (348.15 K). Contrary to the results for GMP/GTP, the error metrics for the CMP/CTP system are not monotonous with temperature. Also, the fitting for this system seems not to be dominated by the error metrics since the enol-tautomer, whose fraction is increasing with temperature, has the largest deviation from the experiment, and the Watson-Crick- and N3-tautomer show nearly the same deviation but strongly differing fractions. Compared to the energy calculations, the fit results are more closely to the cytidine and CMP than to the CTP results.

The last system investigated is the UMP/UTP system. The fit for this system results in the 2-enol-tautomer being the main species with a fraction of 0.686 at 278.15 K, increasing with temperature up to a fraction of 0.900 at 348.15 K. The Watson-Crick-tautomer, the second most abundant form at low temperatures with a nearly 30% population, is vanishing at high temperatures, and the 4-enol-tautomer becomes more populated with populations up to 10 %. These results are mostly not in line with the energetics analysis, but since the fit is not reliable due to the big influence of the phosphate CSA on the calculated chemical shifts of the monophosphates (which seems to accidentally lead to good correlations between some of the experimental nitrogen chemical shifts and some calculated ones with a wrong protonation pattern), the energetics should be trusted. Interesting is the difference in the energetics between the, very similar, UMP/UTP and the TMP/dTTP systems, since for the latter, the Watson-Crick tautomer is calculated to be the dominant species which is in agreement with the UMP, but the UTP main tautomer is the 4-enol-tautomer.

The situation for the nucleoside mono- and triphosphates is more complicated than for the unphosphorylated nucleosides. The energy calculations of the monophosphates result in the Watson-Crick-tautomer being the dominant form for all species excluding GMP, while for the triphosphates GTP is the only one where this tautomer is the most populated form. It is important to consider that the alternative hydrogen bonding-patterns presented by the nucleoside mono- and triphosphates are calculated in water at infinite dilution and are not representative for the situation in a nucleic acid. Within a nucleic acid, the tautomers cannot be stabilized by interactions between base and phosphate chain in a way they can in free solution, for example intramolecular base-phosphate hydrogen bonds cannot be formed. Besides, the conformational freedom is limited due to π-stacking interactions and hydrogen-bond formation with the opposing bases, which should reduce the occurrence of *β*-D-ribose-configurations and *syn*-conformations of the nucleobases. For a complete picture of the tautomeric situation of the nucleotides, a consideration of all possible ionization states would be helpful and possible with the workflow presented here. In this work, only the fully deprotonated species are considered, this way the problem of the determination of the position of the hydrogen atoms within the phosphate chain could be circumvented.

The NMR analysis faced multiple problems: the experiment was done for the triphosphates for which calculations were not possible due to the limits of the computational resources, so the comparison had to be done to calculated shifts of the monophosphates. To overcome this issue, benchmarking of NMR calculations at the DFT level would be necessary to reach the accuracy of the MP2 calculations, also the calibration of reference shielding constants is needed for this. Additionally, the calculated chemical shifts are heavily influenced by the chemical shift anisotropy of the phosphate, leading to low correlations between experiment and calculation and questioning the usefulness of calculations of the triphosphates, since the chemical shifts may be even more influenced by the anisotropy of the phosphate when increasing the chain length. This causes inconsistencies between energy calculations, the correlation between NMR experiment, calculation, and the NMR fit. To solve these problems, a computational approach to take the chemical shift anisotropy into account needs to be developed, since in this work, only the isotropic shielding constants of all species are considered.

The least problematic nucleic acid building blocks investigated in this work are the nucleobases and the nucleosides. Even though the nucleosides have a broad conformational ensemble, which has to be accounted for, they are uncharged since they are lacking a phosphate chain. For the nucleobases and nucleosides, a consistent picture is obtained. The Watson-Crick tautomers are the dominant species. This is shown by energy and NMR calculations as well as an NMR fit. These species are also temperature and pressure (only investigated for the nucleobases) stable. The comparison with data from literature confirmed these results. More problematic are the nucleotide mono- and triphosphates. The EC-RISM results show that the Watson-Crick tautomers are the dominant species for all monophosphates excluding GMP. GMP shows a mixture of the Watson-Crick and the enol-tautomer, this is possible because, in the enol-tautomer main conformation, the sugar conformation is switched in a way that the base is pointing away from the sugar, allowing for favorable hydrogen-bond interactions between the sugar and the phosphate chain. The triphosphates were found to exist in a mixture of tautomers, only the N1-ATP and the N3-CTP are dominant species. This is plausible since the triphosphate chain allows for a multitude of favorable interactions between the phosphates, base and sugar, stabilizing rare tautomers. Unfortunately, the NMR spectroscopy could not help to clarify this situation as expected from the nucleoside results because the CSA of the phosphate heavily influenced the experimental chemical shifts. Another problem is the comparison of the nucleotide results with literature data, since to the best of the author's knowledge very few tautomer fractions of nucleotides in free solution were reported, especially under extreme conditions. Except for the problems faced with the phosphorylated species, the combined computational and NMR spectroscopic approach performed well, yielding results consistently in energetics, comparison with the NMR, and the NMR fit not only for the test-system histamine, also for the unphosphorylated nucleosides.

V. Summary

The goal of this work was the development and testing of a workflow for tautomer prediction in solution, especially under extreme conditions. To achieve this goal, several steps were necessary. To improve the predictivity of the EC-RISM solvation model, the level of theory for the future work had to be chosen. As it should be possible to calculate NMR parameters with high accuracy, the choice fell on MP2/6-311+G(d,p). A PMV correction was parametrized at this level of theory and used within the SAMPL6 and SAMPL7 blind prediction challenges and was extensively benchmarked on the SAMPL2 dataset, which includes experimental reaction free energies for tautomer pairs. To complement the PMV correction for high-pressures, a correction for various temperature conditions was developed and benchmarked using a test set of molecules with known temperature dependent free energies of solvation.

Afterwards, the test system histamine, with multiple degrees of freedom, was studied especially with respect to NMR calculations. This system was used to clarify the nuclei most sensitive for tautomer shifts and to develop a combined computational and NMR spectroscopic approach for tautomer predictions. Since extreme environmental conditions play an important role in this work, a framework for the calculation of NMR parameters at these conditions had to be developed. All these methodological advances and the lessons learned from benchmarking were finally used for the investigation of the tautomer stability of the genetic code at extreme conditions, which likely existed at the early stages of life on the earth, and may play an important role for extraterrestrial life. Not only natural species were investigated but also an expanded genetic alphabet, the Hachimoji code, was explored. The considered nucleic acid building blocks were inspected in an order of increasing complexity; from nucleobases to nucleosides and finally nucleotides. The final goal was reaching consistency between energy calculations and NMR spectroscopy to obtain reliable tautomer fractions for all species at all environmental conditions studied.

Calculations on the test set could show that the PMV correction for temperature variations improves the predictivity of EC-RISM for solvation free energies at various temperatures. With the new level of theory, the tautomer pairs of the SAMPL2 dataset could be calculated with enhanced consistency between different ring sizes of the molecules. Especially the introduction of coupled-cluster calculations, which allow the incorporation of electron correlation effects at a high-level, into the EC-RISM framework, yielded good results. The histamine analysis revealed a problematic atomtype, the n3 type, within the GAFF force field used during 3D RISM iterations. A workflow for the reparametrization of FF parameters using EC-RISM was developed, and with the new n3 parameter, the calculation of histamine pK_a-values is improved significantly. With the combined computational and NMR spectroscopic approach, all possible states of histamine could be elucidated consistently between energetics and NMR spectroscopy. The ^{15}N nucleus could be identified as the most relevant for tautomer predictions of nitrogen heterocycles. In agreement with the IUPAC recommendations, DSS and ammonia were used as reference substances for the calculation of NMR parameters. The reference shielding constants were calculated pressure and temperature dependently, and successfully tested for the prediction of the chemical shifts of the benchmark molecules NMA and TMAO, for which the (pressure dependent) chemical shifts are experimentally known. Multiple referencing methods were tested (a direct, an environment independent, and a newly developed, indirect, referencing method, mimicking the use of high pressure/temperature dependent Ξ-factors). Besides, the pressure and temperature dependent *cis*/*trans* equilibrium of the NMA could be clarified; the increasing *cis* fractions may play an important role in the pressure and temperature unfolding of proteins.

The refined EC-RISM workflow for extreme conditions was successfully applied to the calculation of the tautomer fractions of all natural and multiple non-natural nucleobases, including all Hachimoji bases. The tautomer stability of the natural nucleobases and most of the Hachimoji bases could be demonstrated. The natural nucleobases nearly exclusively present the Watson-Crick hydrogen bonding pattern in aqueous solution, regardless of the environmental conditions, while the Z and especially the B of the Hachimoji code are problematic. High-pressure conditions and modifications in the purine backbone of the B can help to stabilize the main conformation, but it does not become as stable as the natural nucleobases.

The analysis of the more complex systems started with the AMP. For this system, the ^{1}H chemical shifts as well as the population of the main conformation, as suggested from NMR spectroscopy, could be reproduced by using only two conformations. These promising results led to investigations of all natural nucleosides, this time considering the whole conformational ensemble. The examination of the tautomeric situation of these species revealed, consistent to literature and the nucleobase results, the stability of the Watson-Crick tautomers over the whole temperature range analyzed. For this and the analysis of the nucleotides, temperature dependent experimental NMR chemical shifts, recorded during an RESOLV internship in the laboratory of Roland Sigel's group at UZH Zürich, were used. The consistency between the nucleobase results, energy calculations, the literature and the NMR, which are achieved for the nucleosides, are promising, so the workflow was applied to the nucleotides. Unfortunately, the calculations did not work out as expected. The main reason for this are problems with the NMR calculations: The NMR workflow was only developed for isotropic chemical shifts, while the spectra of the nucleotides are heavily influenced by the chemical shift anisotropy of the phosphate chain. Therefore, further investigations of anisotropic effects on the calculation of NMR chemical shifts are needed. Additionally, the phosphate chain of nucleoside mono- and especially triphosphates allows many favorable interactions with base and sugar, which helps to stabilize rare tautomers in free solution (which is not possible for nucleobases and nucleosides), and the comparison to external results was complicated due to a lack of literature data. Apart from these issues with the phosphorylated species, the combined computational and NMR spectroscopic workflow proved to be reliable for the prediction of tautomer fractions of flexible molecules even under extreme conditions.

One of the most important results of the work is that the stability of all nucleic acid building blocks is influenced by the environmental conditions. However, the natural species are already so stable that their stability in the studied range is only slightly influenced. This is not the case for the non-natural species, like the Hachimoji nucleobases Z and B, which are already unstable at normal conditions and are decisively influenced by the small energy change. The high reaction free energy baseline of the natural species ensures that their stability is given over a wide range of environmental conditions, and they are probably also well suited for early life conditions and extraterrestrial conditions. They seem to be robust and universally applicable in a large part of the universe. The non-natural species, on the other hand, need certain conditions to be stable. They can probably be used in life forms, however, where the respective suitable environmental conditions are present. In order to be able to predict these conditions, not only the influence of temperature and pressure on the tautomerization process must be investigated, as has been done in this work, but also the simultaneous influence of both parameters, and the influence of other parameters expected at early life conditions, like high concentrations of electrolytes, must be checked.

In principle, different possibilities are conceivable for this. On the one hand, the two PMV corrections, for temperature and pressure, can be applied together, an approach that would require not only the calculation of new, pressure and temperature dependent, solvent susceptibilities and a sufficient sampling of points in the *p*/*T*-diagram, but also extensive benchmarking using experimental data that are difficult to access. On the other hand, there are two promising alternative approaches, which could be used for combined pressure and temperature dependent calculations. At first, the pressure dependency of the energetics in solution can be considered via the integration of the PMV, which is, as a short introduction to the method, shown in Eqn.

165 and 166. The author's first attempts with this methodology were promising, and the workgroup is currently working with it on high pressure phenomena. Secondly, temperature dependent calculations can be used to approximate the reaction enthalpy ΔH via the (integrated) van’t Hoff equation. The pressure and temperature dependent fractions of a minor tautomer can afterwards be calculated with knowledge of the reaction free energy at ambient conditions by a partition function approach assuming the pressure, respectively temperature independence of the PMV and the reaction enthalpy. Due to the still high uncertainty of these absolute quantities it remains to be seen whether such an approach provides useful extrapolations. This way, only the PMV correction for ambient conditions would be needed for the calculation of the whole p/T-dependent stability diagram of the nucleic acid building blocks. This could not only lead to answers to open questions about the universality and robustness of the genetic code in the universe, but also help synthetic biology to develop non-natural nucleic acid building blocks that have the desired tautomeric properties for particular environmental conditions. These perspectives are currently investigated further.

References

1 F. Crick, *Nature* 227, 5258 (1970)

2 W. Martin, M. J. Russell, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358, 1429, 59 (2003)

3 S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H.-J. Kim, A. M. Bates, N. E. Watkins Jr., H. A. SantaLucia, A. J. Meyer, S. DasGupta, J. A. Piccirilli, A. D. Ellington, J. SantaLucia Jr., M. M. Georgiadis, S. A. Benner, *Science*, 363, 884 (2019)

4 N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des*. 32, 1151 (2018)

5 N. Tielker, L. Eberlein, C. Chodun, S. Güssregen, S. M. Kast, *J. Mol. Model*. 25, 139 (2019)

6 C. E. Munte, M. Karl, W. Kauter, L. Eberlein, T.-V. Pham, M. Beck Erlach, S. M.Kast, W. Kremer, H. R.Kalbizer, *Biophys. Chem*. 254, 106261 (2019)

7 T. Pongratz, P. Kibies, L. Eberlein, N. Tielker, C. Hölzl, S. Imoto, M. Beck Erlach, S. Kurrmann, P. H. Schummel, M. Hofmann, O. Reiser, R. Winter, W. Kremer, H. R. Kalbitzer, D. Marx, D. Horinek, S. M. Kast, *Biophys. Chem.* 257, 106258 (2020)

8 N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des*. 34, 453 (2020)

9 L. Eberlein, F. R. Beierlein, N. J. R. van Eikema Hommes, A. Radadiya, J. Heil, S. A. Benner, T. Clark, S. M. Kast, N. G. J. Richards, *J. Chem. Theory Comput*. 16, 4, 2766 (2020)

10 N. Tielker, L. Eberlein, G. Hessler, K. F. Schmidt, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des*. https://doi.org/10.1007/s10822-020-00347-5 (2020)

11 C. Laar, *Chem. Ber*. 18, 648 (1885)

12 C. Laar, *Chem. Ber*. 19, 730 (1886)

13 W. Marckwald, *Liebigs Ann. Chem.* 286, 343 (1895)

14 P. Müller, *Pure Appl. Chem*. 66, 1077 (1994)

15 S. Alonso, N. Izagirre, I. Smith-Zubiaga, J. Gardeazabal, J. L. Díaz-Ramón, J. L. Díaz-Pérez, D. Zelenika, M. D. Boyano, N. Smit, C. de la Rúa, *BMC Evolutionary Biology*, 8, 74 (2008)

16 Z. Jiao, Z. G. Zhang, T. J. Hornyak, A. Hozeska, R. L. Zhang, Y. Wang, L. Wang, C. Roberts, F. M. Strickland, M. Chopp, *Developmental Biology*, 296, 396 (2006)

17 X. Yan, T. Hollis, M. Svinth, P. Day, A. F. Monzingo, G. W. Milne, J. D. Robertus, *J. Mol. Biol*. 266, 1043 (1997)

18 J. Heil, „Effiziente Fluidphasentheorie für Protonierungsprozesse in komplexen Systemen“, *Dissertation*, Technische Universität Dortmund, 2016.

19 L. Antonov, V. Deneva, S. Simeonov, V. Kurteva, D. Nedeltcheva, J. Wirz, *Angew. Chem. Int. Ed*. 48, 7875 (2009)

20 S. A. Tawfik, X. Y. Cui, S. P. Ringer, C. Stampfl, *J. Chem. Theory Comput*. 11, 9, 4154 (2015)

21 A. R. Katritzky, C. D. Hall, B. E. M. El-Gendy, B. Draghici, *J. Comput.-Aided Mol. Des*. 24, 475 (2010)

22 W. Auwärter, K. Seufert, F. Bischoff, D. Ecija, S. Vijayaraghavan, S. Joshi, F. Klappenberger, N. Samudrala, J. V. Barth, *Nature Nanotechnology*, 7, 41 (2012)

23 T. Kumagai, L. Grill, “Direct Observation and Control of Single-Molecule Tautomerization by Low-Temperature Scanning Tunneling Microscopy” in *Tautomerism Concepts and Applications in Science and Technology*, (ed. L. Antonov) Wiley, 162 (2016)

24 B. Bax, C. Chung, C. Edge, *Acta Cryst*. 73, 131 (2017)

25 J. Zhou, O. Kostko, C. Nicolas, X. Tang, L. Belau, M. S. de Vries, M. Ahmed, *J. Phys. Chem. A*, 113, 17 (2009)

26 J. Wirz, “Kinetic studies of keto-enol and other tautomeric equilibria by flash photolysis” in *Advances in Physical Organic Chemistry*, 44 (ed. J.P. Richard), Academic Press, 325 (2010)

27 L. Antonov (ed.), *Tautomerism Methods and Theories*, Wiley (2013)

28 L. Antonov, “Absorption UV–vis Spectroscopy and Chemometrics: From Qualitative Conclusions to Quantitative Analysis” in *Tautomerism Methods and Theories*, (ed. L. Antonov) Wiley (2013)

29 R.M. Claramunt, C. López, M. D. Santa Maŕia, D. Sanz, J. Elguero, *Prog. Nucl. Mag. Res. Sp.* 49, 169 (2006)

30 Y. Manolova, V. Deneva, L. Antonov, E. Drakalska, D. Momekova, N. Lambov, *Spectrochim. Acta A*, 132, 815 (2014)

31 A. Mazzanti, D. Casarini, *WIREs Comput Mol. Sci.* 2, 613 (2012)

32 H. Friebolin, *Ein-und zweidimensionale NMR-Spektroskopie: eine Einführung*; Wiley, (2013)

33 L. M. Nguyen, J. Roche, *J. Magn. Reson*. 277, 179 (2017)

34 J. Roche, C. A. Royer, C. Roumestand, *Prog. Nucl. Mag. Res. Sp.* 102-103, 15 (2017)

35 W. A. Thomas, *Prog. Nucl. Mag. Res. Sp.* 30, 3-4, 183 (1997)

36 M. Karplus, *J. Chem. Phys.* 30, 1, 11 (1959)

37 M. Karplus, *J. Am. Chem. Soc.* 85, 18, 2870 (1963)

38 C. Thibaudeau, R. Stenutz, B. Hertz, T. Klepach, S. Zhao, Q. Wu, I. Carmichael, A. S. Serianni, *J. Am. Chem. Soc.* 126, 48, 15668 (2004)

39 G. Bifulco, P. Dambruoso, L. Gomez-Paloma, R. Riccio, *Chem. Rev.* 107, 9, 3744 (2007)

40 C. M. Thiele, *Eur. J. Org. Chem.* 34, 5673 (2008)

41 J. M. Seco, E. Quiñoá, R. Riguera, *Chem. Rev.* 104, 1, 17 (2004)

42 N. Matsumori, D. Kaneno, M. Murata, H. Nakamura, K. Tachibana, K. *J. Org. Chem.* 64, 3, 866 (1999)

43 E. Kleinpeter, “NMR Spectroscopic Study of Tautomerism in Solution and in the Solid State” in *Tautomerism Methods and Theories*, (ed. L. Antonov) Wiley (2013)

44 W. F. Reynolds, C. W. Tzeng, *Can. Jo. Biochem.* 55, 5, 576 (1977)

45 L. Antonov (ed.), *Tautomerism Concepts and Applications in Science and Technology*, Wiley (2016)

46 R. Ditchfield, *Mol. Phys.* 27, 4, 789 (1974)

47 M. Schindler, W. Kutzelnigg, *J. Chem. Phys.* 76, 4, 1919 (1982)

48 M. Dračínský, H. M. Möller, T. E. Exner, *J. Chem. Theory Comput.* 9, 8, 3806 (2013)

49 C. I. Nieto, P. Cabildo, M. Á. García, R. M. Claramunt, I. Alkorta, J. Elguero, *Beilstein J. Org. Chem.* 10, 1620 (2014)

50 E. Kleinpeter, L. Hilfert, A. Koch, *J. Phys. Org. Chem.* 13, 473 (2000)

51 F. B. Miguel, J. A. Dantas, S. Amorim, G. F. S. Andrade, L. A. S. Costa, M. R. C. Couri, *Spectrochim. Acta A* 152, 318 (2016)

52 A. Contini, D. Nava, P. Trimarco, *J. Org. Chem.* 71, 159 (2006)

53 A. Skotnicka, P. Czeleń, R. Gawinecki, *J. Mol. Struct.* 1134, 546, (2017)

54 M. G. Siskos, P. C. Varras, I. P. Gerothanassis, *Tetrahedron* 76, 9, 130979 (2020)

55 R. B. Campos, L. R. A. Menezes, A. Barison, D. J. Tantillo, E. S. Orth, *Chem. Eur. J.* 22, 43, 15521 (2016)

56 M. Flores-Leonar, N. Esturau-Escofet, J. M. Méndez-Stivalet, A. Marı́n-Becerra, C. Amador-Bedolla, *J. Mol. Struct*. 1006, 600 (2011)

57 O. Cala, F. Guillière, I. Krimm, *Anal. Bioanal. Chem*. 406, 943 (2014)

58 B. Meyer, T. Peters, *Angew. Chem. Int. Ed*. 42, 8 (2003)

59 K. Farah, F. Müller-Plathe, M. C. Böhm, *ChemPhysChem*, 13, 5, 1127 (2012)

60 A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, *J. Phys. Chem. A*, 105, 9396 (2001)

61 J. Huang, M. Meuwly, "Force Field Treatment of Proton and Hydrogen Transfer in Molecular Systems" in *Tautomerism Methods and Theories*, (ed. L. Antonov) Wiley (2013)

62 V. Fock, *Z. Phys*. 61, 126 (1930)

63 C. Møller, M. S. Plesset, *Phys. Rev*. 46, 618 (1934)

64 F. Coester, *Nucl. Phys*. 7, 421 (1958)

65 O. Sinanoğlu, *J. Chem. Phys*. 36, 706 (1962)

66 R. J. Bartlett, M. Musiał, *Rev. Mod. Phys*. 79, 291 (2007)

67 P. Hohenberg, W. Kohn, *Phys. Rev*. 136, 864 (1964)

68 W. Kohn, L. J. Sham, *Phys. Rev*. 140, 1133 (1965)

69 J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.*, 77, 3865 (1996)

70 J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.*, 78 ,1396 (1997)

71 A. D. Becke, *J. Chem. Phys*. 98, 5648 (1993)

72 C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* 37, 785 (1988)

73 S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys*. 58, 1200 (1980)

74 S. Grimme, *J. Chem. Phys*. 124, 034108 (2006)

75 J. S. Dewar, E. G. Zoebisch, E. F. Healy, *J. Am. Chem. Soc*. 107, 3902 (1985)

76 J. J. P. Stewart, *J. Mol. Model*. 13, 1173 (2007)

77 R. Car, M. Parrinello, *Phys. Rev. Lett*. 55, 2471 (1985)

78 D. Marx, J. Hutter, "Ab initio molecular dynamics: Theory and Implementation" in *Modern Methods and Algorithms of Quantum Chemistry*, J. Grotendorst (Ed.), John von Neumann Institute for Computing (2000)

79 S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* 132, 15, 154104 (2010)

80 S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.* 32, 7, 1456 (2011)

81 A. D. Becke, E. R. Johnson, *J. Chem. Phys.* 123, 15, 154101 (2005)

82 J. Engel, A. Richters, M. Getlik, S. Tomassi, M. Keul, M. Termathe, J. Lategahn, C. Becker, S. Mayer-Wrangowski, C. Grütter, N. Uhlenbrock, J. Krüll, N. Schaumann, S. Eppmann, P. Kibies, F. Hoffgaard, J. Heil, S. Menninger, S. Ortiz-Cuaran, J. M. Heuckmann, V. Tinnefeld, R. P. Zahedi, M. L. Sos, C. Schultz-Fademrecht, R. K. Thomas, S. M. Kast, D. Rauh, *J. Med. Chem.* 58, 17, 6844 (2015)

83 M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chem. Rev.* 112, 3, 1839 (2012)

84 J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* 105, 8, 2999 (2005)

85 D. A. Zichi, P. J. Rossky, *J. Chem. Phys.* 84, 3, 1712 (1986)

86 M. Dračínský, P. Bouř, *J. Chem. Theory Comput.* 6, 1, 288 (2010)

87 T. Yamazaki, H. Sato, F. Hirata, *Chem. Phys. Lett.* 325, 668 (2000)

88 T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B* 112, 14, 4337 (2008)

89 R. Frach, S. M. Kast, *J. Phys. Chem. A* 118, 49, 11620 (2014)

90 M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, P. J. Taylor, *J. Comput.-Aided Mol. Des.* 24, 259 (2010)

91 S. M. Kast, J. Heil, S. Güssregen, K. F. Schmidt, *J. Comput.-Aided Mol. Des*. 24, 343 (2010)

92 M. J. Nowak, L. Lapinski, J. Fulara, A. Les, L. Adamowicz, *J. Phys. Chem*. 96, 1562 (1992)

93 L. D. Hatherley, R. D. Brown, P. D. Godfrey, A. P. Pierlot, W. Caminad, D. Damiani, S. Melandri, L. B. Pavero, *J. Phys. Chem*. 97, 46 (1993)

94 U. Ohms, H. Guth, E. Hellner, H. Dannöhl, A. Schweig, *Z. Kristallogr. – Cryst. Mater*. 169, 185 (1984)

95 H. W. Yang, B. M. Craven, *Acta Cryst*. 54, 912 (1998)

96 J. Frank, A. R. Katritzky, *J. Chem. Soc. Perkin Trans*. 2, 12, 1428 (1976)

97 L. Forlani, G. Cristoni, C. Boga, P. E. Todesco, E. Del Vecchio, S. Selva, M. Monari. *Arkivoc*, 11, 198 (2002)

98 G. G. Hammes, P. J. Lillford, *J. Am. Chem. Soc*. 92, 26, 7578 (1970)

99 J. M. Rawson, R. E.P. Winpenny, *Coord. Chem. Rev*. 139, 313 (1995)

100 A. Klamt, M. Diedenhofen, *J. Comput.-Aided Mol. Des*. 24, 621 (2010)

101 P. Khuu, P. S. Ho, *Biochemistry*, 48, 7824 (2009)

102 W. Wang, H. W. Hellinga, L. S. Beese, *Proc. Natl. Acad. Sci. U.S.A*. 108, 43, 17644 (2011)

103 V. H. Harris, C- L- Smith, W. J. Cummins, A. L. Hamilton, H. Adams, M. Dickman, D. P. Hornby, D. M. Williams, *J. Mol. Biol.* 326, 1389 (2003)

104 V. Singh, B. I. Fedeles, J. M. Essigmann, *RNA*, 21, 1 (2014)

105 M. Piacenca, S. Grimme, *J. Comput. Chem*. 25, 83 (2004)

106 M. K. Shukla, J. Leszczynski, *WIREs Comput. Mol. Sci*. 3, 637 (2013)

107 M. S. de Vries, “Tautomer-Selective Spectroscopy of Nucleobases, Isolated in the Gas Phase” in *Tautomerism Methods and Theories*, (ed. L. Antonov) Wiley (2013)

108 S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H.-J. Kim, A. M. Bates, N. E. Watkins Jr., H. A. SantaLucia, A. J. Meyer, S. DasGupta, J. A. Piccirilli, A. D. Ellington, J. SantaLucia Jr., M. M. Georgiadis, S. A. Benner, *Science*, 363, 884 (2019)

109 P. W. Atkins, J. de Paula, *Physikalische chemie* Wiley (2013)

110 M. Hesse, H. Meier, B. Zeeh, *Spektroskopische Methoden in der organischen Chemie* Thieme Verlag (1991)

111 R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, R. Goodfellow, P. Granger, *Pure Appl. Chem.* 73, 11, 1795 (2001)

112 P. W. Atkins, R. S. Friedman, *Molecular quantum mechanics* Oxford university press (2011)

113 M. Born, R. Oppenheimer, *Ann. Phys*. 20, 457 (1927)

114 A. Szabo, N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation (2012)

115 I. N. Levine, *Quantum Chemistry* Pearson (2014)

116 J. L. Whitten, *J. Chem. Phys*. 58, 4496 (1973)

117 B. I. Dunlap, J. W. D. Connolly, J. R. Sabin, *J. Chem. Phys*. 71, 3396 (1979)

118 O. Vahtras, J. Almlöf, M. W. Feyereisen, *Chem. Phys. Lett*. 213, 5-6, 514 (1993)

119 F. Neese, *J. Comput. Chem*. 24, 14, 1740 (2003)

120 A. M. Burow, M. Sierka, F. Mohamed, *J. Chem. Phys*. 131, 214101 (2009)

121 K. Eichkorn, F. Weigend, O. Treutler, R. Ahlrichs, *Theor. Chem. Acc*. 97, 119 (1997)

122 K. Eichkorn, O. Treutler, H. Öhm, M. Häser, R. Ahlrichs, *Chem. Phys. Lett*. 240, 283 (1995)

123 F. Weigend, M. Haser, H. Patzelt, R. Ahlrichs, *Chem. Phys. Lett*. 294, 143 (1998)

124 F. Weigend, *Phys. Chem. Chem. Phys*. 8, 1057 (2006)

125 W. Kutzelnigg, *Theor. Chim. Acta*. 68, 445 (1985)

126 S. Ten-no, *Chem. Phys. Lett*. 398, 56 (2004)

127 D. P. Tew, W. Klopper, C. Neiss, C. Hättig, *Phys. Chem. Chem. Phys*. 9, 1921 (2007)

128 E. F. Valeev, Chem. Phys. Lett. 395, 190 (2004)

129 F. Pavošević, P. Pinski, C. Riplinger, F. Neese, E. F. Valeev, *J. Chem. Phys*. 144, 144109 (2016)

130 F. Neese, *Interdisc. Rev. Comput. Mol. Sci.* 2, 73 (2012)

131 A. D. Becke, *Phys. Rev. A* 38, 3098 (1988)

132 C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* 37, 785 (1988)

133 P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* 98, 45, 11623 (1994)

134 S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* 58, 1200 (1980)

135 N. F. Ramsey, *Phys. Rev.* 78, 6, 699 (1950)

136 P. Pyykko, Theor. Chem. Acc. 103, 214 (2000)

137 T. Helgaker, M. Jaszuński, K. Ruud, *Chem. Rev.* 99, 1, 293 (1999)

138 J. C. Facelli, *Concepts Magn. Reson.* 20A, 1, 42 (2004)

139 T. Helgaker, S. Coriani, P. Jørgensen, K. Kristensen, J. Olsen, K. Ruud, *Chem. Rev.* 112, 1, 543 (2012)

140 J. Vaara, *Phys. Chem. Chem. Phys.* 9, 40, 5399 (2007)

141 I. L. Rusakova, L. B. Krivdin, Y. Y. Rusakov, A. B. Trofimov, *J. Chem. Phys.* 137, 4, 44119 (2012)

142 Y. Y. Rusakov, L. B. Krivdin, *Russ. Chem. Rev.* 82, 2, 99 (2013)

143 R. M. Stevens, R. M. Pitzer, W. N. Lipscomb, *J. Chem. Phys.* 38, 2, 550 (1963)

144 A. E. Hansen, T. D. Bouman, *J. Chem. Phys.* 82, 11, 5035 (1985)

145 A. E. Hansen, A. E.; Bouman, *J. Chem. Phys.* 91, 6, 3552 (1989)

146 T. A. Keith, R. F. Bader, *Chem. Phys. Lett.* 194, 1-2, 1 (1992)

147 T. A. Keith, R. F. Bader, *Chem. Phys. Lett.* 210, 1-3, 223 (1993)

148 J. R. Cheeseman, G. W. Trucks, T. A. Keith, M. J. Frisch, *J. Chem. Phys.* 104, 14, 5497 (1996)

149 R. Ditchfield, *Mol. Phys.* 27, 4, 789 (1974)

150 J. Pople, *Faraday Discuss.* 34, 7 (1962)

151 J. Gauss, J. F. Stanton, *J. Chem. Phys*. 102, 1 (1995)

152 F. London, *J. Phys. Radium* 8, 10, 397 (1937)

153 J. Gauss, J. F. Stanton, “II. Basic aspects for the calculation of NMR chemical shifts” in *Advances in chemical physics Vol. 123*, I. Prigogine, S. A. Rice (Ed.), Wiley (2002)

154 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, *Gaussian 09*, Revision E 01, Gaussian Inc. Wallingford CT (2009)

155 K. Wolinski, J. F. Hinton, P. Pulay, *J. Am. Chem. Soc.* 112, 23, 8251 (1990)

156 T. Ziegler, G. Schreckenbach, *J. Phys. Chem.* 99, 606 (1995)

157 S. K. Wolff, T. Ziegler, *J. Chem. Phys.* 109, 3, 895 (1998)

158 D. Flaig, M. Maurer, M. Hanni, K. Braunger, L. Kick, M. Thubauville, C. Ochsenfeld, *J. Chem. Theory Comput.* 10, 2, 572 (2014)

159 B. Nagy, F. Jensen, "Basis sets in quantum chemistry" in *Reviews in computational chemistry Vol. 30*, A. L. Parrill, K. B. Lipkowitz (Ed.) Wiley (2017)

160 J. C. Slater, *Phys. Rev.* 36, 57 (1930)

161 S. Huzinaga, *Comput. Phys. Rep.* 2, 279 (1985)

162 J. K. Lebanowski, "Simplified introduction to ab initio basis sets. Terms and notation" (1996) from www.ccl.net/cca/documents/basis-sets/basis.html (accessed March 31, 2020).

163 R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.* 72, 1, 650 (1980)

164 T. H. Dunning, *J. Chem. Phys.* 90, 2, 1007 (1989)

165 R. A. Kendall, T. H. Dunning, R. J. Harrison, *J. Chem. Phys.* 96, 9, 6796 (1992)

166 D. E. Woon, T. H. Dunning, *J. Chem. Phys*, 103, 11, 4572 (1995)

167 F. Jensen, *J. Chem. Phys.* 115, 20, 9113 (2001)

168 F. Jensen, *J. Chem. Phys.* 116, 17, 7372 (2002)

169 F. Jensen, *J. Chem. Phys.* 117, 20, 9234 (2002)

170 F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys*. 7, 3297 (2005)

171 F. Weigend, *Phys. Chem. Chem. Phys*. 8, 1057 (2006)

172 J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, T. Head-Gordon, *J. Phys. Chem. B* 114, 2549 (2010)

173 T. P Senftle, S. Hong, M. M. Islam, S. B Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, A. C. T. van Duin, *Npj Comput. Mater*. 15011 (2016)

174 W. F. van Gunsteren, H. J. C. Berendsen, *Angew. Chem.* 102, 1020 (1990)

175 W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, H. B. Yu, *Angew. Chem.* 118, 25, 4168 (2006)

176 A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard III, W. M. Skiff, *J. Am. Chem. Soc.* 114, 25, 10024 (1992)

177 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comp. Chem.* 4, 187

(1983)

178 B. R. Brooks, C. L. Brooks III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comp. Chem*. 30, 1545 (2009)

179 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc*. 117, 5179 (1995)

180 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput*. 11, 3696 (2015)

181 C. Oostenbrink, A. Villa, A. E. Mark, W. F. van Gunsteren, *J. Comput. Chem.* 25, 13, 1656 (2004)

182 W. L. Jorgensen, J. Tirado-Rives, *J. Am. Chem. Soc*. 110, 6, 1657 (1988)

183 T. A. Halgren, *J. Comput. Chem*. 20, 7, 720 (1999)

184 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* 25, 9, 1157 (2004)

185 J. Barzilai, J. M. Borwein, *IMA J. Numer. Anal*. 8, 1, 141 (1988)

186 Avogadro: an open-source molecular builder and visualization tool. Version 1.20, http://avogadro.cc/

187 L. Verlet, *Phys. Rev.* 159, 1, 98 (1967)

188 P. Ewald, *Ann. Phys*. 369, 3, 253 (1921)

189 T. Darden, L. Perera, L. Li, L. Pedersen, *Structure* 7, 3, 55 (1999)

190 D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon- Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York, P.A. Kollman, AMBER 2018, University of California, San Francisco (2018)

191 J. G. Kirkwood, *J. Chem. Phys.* 3,5, 300 (1935)

192 B. Roux, T. Simonson, *Biophys. Chem.* 78, 1-2, 1 (1999)

193 S. Decherchi, M. Masetti, I. Vyalov, W. Rocchia, *Eur. J. Med. Chem*. 91, 27 (2015)

194 J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, K. Schulten, *J. Comp. Chem*. 26, 1781 (2005)

195 J. Tomasi, M. Persico, *Chem. Rev*. 94, 2027 (1994)

196 G. Scalmani, M. J. Frisch, *J. Chem. Phys.* 132, 11, 114110 (2010)

197 J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* 105, 2999 (2005).

198 S. Miertus, E. Scrocco, J. Tomasi, *Chem. Phys.* 55, 117 (1981)

199 V. Barone, M. Cossi, *J. Phys. Chem. A* 102, 11, 1995 (1998)

200 E. Cancès, B. Mennucci, J. Tomasi, *J. Chem. Phys.* 107, 8, 3032 (1997)

201 A. Klamt, G. Schüürmann, *J. Chem. Soc. Perkin Trans.* 2, 799 (1993)

202 R. Cammi, B. Mennucci, J. Tomasi, *J. Chem. Phys.* 110, 16, 7627 (1999)

203 G. Sigalov, P. Scheffel, A. Onufriev, *J. Chem. Phys*. 122, 094511 (2005)

204 G. Sigalov, A. Fenley, A. Onufriev, *J. Chem. Phys*. 124, 124902 (2006)

205 J.-P. Hansen, I. R. McDonald, *Theory of simple liquids*, 3rd Ed, Elsevier (2007)

206 R. Evans, *Adv. Phys*. 28, 2, 143 (1979)

207 T. Morita, K. Hiroike, *Prog. theor. Phys*. 25, 537 (1961)

208 L. Blum, A. J. Torruella, *J. Chem. Phys*. 56, 303 (1972)

209 H. C. Andersen, D. Chandler, *J. Chem. Phys. 57*, 1918 (1972)

210 D. Chandler, H. C. Andersen, *J. Chem. Phys*. 57, 1930 (1972)

211 F. Hirata, P. J. Rossky, B. Montgomery Pettitt, *J. Chem. Phys*. 78, 4133 (1983)

212 E. L. Ratkova, D. S. Palmer, M. V. Fedorov, *Chem. Rev*. 115, 6312 (2015)

213 F. Hirata (Ed.), *Molecular theory of solvation* Springer (2003)

214 S. M. Kast, *Phys. Rev. E* 67, 4 ,1, 41203 (2003)

215 A. Kovalenko, F. Hirata, *J. Chem. Phys*. 113, 2793 (2000)

216 S. M. Kast, T. Kloss, *J. Chem. Phys*. 129, 236101 (2008)

217 F. Hirata, P. J. Rossky, *Chem. Phys. Lett*. 83, 2, 329 (1981)

218 J. Perkyns, B. Montgomery Pettitt, *Chem. Phys. Lett.* 190, 626 (1992)

219 J. Perkyns, B. Montgomery Pettitt, *J. Chem. Phys.* 97, 7656 (1992)

220 D. Beglov, B. Roux, *J. Phys. Chem. B* 101, 7821 (1997)

221 A. Kovalenko, F. Hirata, *Chem. Phys. Lett.* 290, 237 (1998)

222 H. A. Lorentz, *Ann. Phys*. 248, 1, 127 (1881)

223 D. Berthelot, *C. R. Acad. Sci*. 126, 1703 (1898)

224 P. Pulay, *Chem. Phys. Lett.* 73, 2, 393 (1980)

225 A. Kovalenko, S. Ten-No, F. Hirata, *J. Comput. Chem.* 20, 9, 928 (1999)

226 P. Ewald, *Ann. Phys*. 369, 253 (1921)

227 J. Heil, S. M. Kast, *J. Chem. Phys*. 142, 114107 (2015)

228 S. Ten-no, F. Hirata, S. Kato, *Chem. Phys. Lett*. 214, 3, 391 (1993)

229 H. Sato, *Phys. Chem. Chem. Phys*. 15, 7450 (2013)

230 M. Reimann, M. Kaupp, *J. Phys. Chem. A* 124, 7439 (2020)

231 F. Hoffgaard, J. Heil, S. M. Kast, *J. Chem. Theory Comput*. 9, 11, 4718 (2013)

232 N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, Stefan M. Kast, *J. Comput.-Aided Mol. Des*. 30, 1035 (2016)

233 J. Heil, D. Tomazic, S. Egbers, S. M. Kast, *J. Mol. Model*. 20, 2161 (2014)

234 R. Frach, P. Kibies, S. Böttcher, T. Pongratz, S. Strohfeldt, S, Kurrmann, J. Koehler, M. Hofmann, W. Kremer, H. R. Kalbitzer, O. Reiser, D. Horinek, S. M. Kast, *Angew. Chem. Int. Ed. 55*, 8757 (2016)

235 V. Sergiievskyi, G. Jeanmairet, M. Levesque, D. Borgis, *J. Chem. Phys*. 143, 184116 (2015)

236 D. S. Palmer, A. I. Frolov, E. L. Ratkova, M. V. Fedorov, *J. Phys. Condens. Matter* 22, 492101 (2010)

237 D. Tomazic, „Optimizing free energy functionals in integral equation theory“, *Dissertation*, Technische Universität Dortmund, 2016.

238 J. G. Kirkwood, F. P. Buff, *J. Chem. Phys*. 19, 774 (1951)

239 T. Imai, M. Kinoshita, F. Hirata, *J. Chem. Phys*. 112, 9469 (2000)

240 P. G. Kusalik, G. N. Patey, *J. Chem. Phys*. 86, 5110 (1987)

241 P. G. Kusalik, G. N. Patey, *J. Chem. Phys*. 89, 5843 (1988)

242 M. Misin, M. V. Fedorov, D. S. Palmer, *J. Phys. Chem. B*, 120, 975 (2016)

243 M. A. Kastenholz, P. H. Hünenberger, *J. Chem. Phys*. 124, 224501 (2006);

244 S. M. Kathmann, I.-F. W. Kuo, C. J. Mundy, *J. Am. Chem. Soc*. 130, 16556 (2008)

245 E. Harder, B. Roux, *J. Chem. Phys*. 129, 234706 (2008)

246 S. M. Kathmann, I-F. W. Kuo, C. J. Mundy, G. K. Schenter, *J. Phys. Chem. B*, 115, 4369 (2011)

247 A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer, D. G. Truhlar, Minnesota Solvation Database – version 2012, University of Minnesota, Minneapolis (2012)

248 R. Cammi, C. Cappelli, B. Mennucci, J. Tomasi, *J. Chem. Phys*. 137, 154112 (2012)

249 R. Cammi, *J. Comput. Chem*. 36, 2246 (2015).

250 J. Johnson, D. A. Case, T. Yamazaki, S. Gusarov, A. Kovalenko, T. Luchko, *J. Phys.: Condens. Matter*, 28, 344002 (2016)

251 M. Misin, M. V. Fedorov, D. S. Palmer, *J. Chem. Phys*. 142, 091105 (2015)

252 A. C. Chamberlin, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B*, 110, 5665 (2006)

253 A. C. Chamberlin, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B*, 112, 3024 (2008)

254 A. Ben-Naim, *J. Chem. Phys*. 82, 792 (1978)

255 M. S. Head, J. A. Given, M. K. Gilson, *J. Phys. Chem. A* 101, 1609 (1997)

256 M. D. Tissandier, K. A. Cowen, A. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, *J. Phys. Chem. A* 102, 7787 (1998)

257 H. Zhang, Y. Jiang, H. Yan, Z. Cui, Y. Chunhua, *J. Chem. Inf. Model*. 57, 2763 (2017)

258 J. J. Klicić, R. A. Friesner, S. Y. Liu, W. C. Guida, *J. Phys. Chem. A* 106, 1327 (2002)

259 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, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 16*, Revision C.01, Gaussian Inc., Wallingford CT, 2016.

260 H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem*. 91, 6269 (1987)

261 S. Maw, H. Sato, S. Ten-no, F. Hirata, *Chem. Phys. Lett*. 276, 20 (1997)

262 W. B. Floriano, M. A. C. Nascimento, *Braz. J. Phys.* 34, 38 (2004)

263 L. E. Chirlian, M. M. Francl, *J. Comput. Chem.* 8, 894 (1987)

264 The International Association for the Properties of Water and Steam, „Release on the Static Dielectric Constant of Ordinary Water Substance for Temperatures from 238 K to 873 K and Pressures up to 1000 MPa", IAPWS

R8-97, Erlangen, Germany (1997)

265 The International Association for the Properties of Water and Steam, „Revised Release on the IAPWS Industrial Formulation 1997 for the Thermodynamic Properties of Water and Steam (The revision only relates to the extension of region 5 to 50 MPa)" IAPWS R7-97(2012), Lucerne, Switzerland (2007)

266 S. M. Kast, private communication

267 L. Martínez, R. Andrade, E. G. Birgin, J. M. Martínez, *J. Comput. Chem*. 30, 13, 2157 (2009)

268 A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem*. 23, 16, 1623 (2002)

269 S. Miyamoto, P. A. Kollman, *J. Comput. Chem*. 13, 952 (1992)

270 T. Darden, D. York, L. Pedersen, *J. Chem. Phys*. 98,10089 (1993)

271 G. J. Martyna, D. J. Tobias, M. L. Klein, *J. Chem. Phys*. 101, 4177 (1994)

272 S. E. Feller, Y. Zhang, R. W. Pastor, B. R. Brooks, *J. Chem. Phys*. 103, 4613 (1995)

273 T. B. Paiva, M. Tominaga, A. C. M. Paiva, *J. Med. Chem.* 13, 986 (1970)

274 J. Auch, „Optimierung von Modellen zur Vorhersage von Freien Hydratationsenthalpien", Bachelor's Thesis, Technische Universität Dortmund, 2018

275 C. Chodun, "Optimierung von nichtbindenden Solvat-Solvens-Wechselwirkungsparametern auf Basis von Freie-Energie-Ableitungen", Master's Thesis, Technische Universität Dortmund, 2020

276 F. Mrugalla, S. M. Kast, *J. Phys.: Condens. Matter*, 28, 344004 (2016)

277 L. Eberlein, „Konformationsaufklärung von pH-abhängigen Histamin-Spezies", *Master's Thesis*, Technische Universität Dortmund, 2016

278 D. Flaig, M. Maurer, M. Hanni, K. Braunger, L. Kick, M. Thubauville, C. Ochsenfeld, *J. Chem. Theory Comput.* 10, 2, 572 (2014)

279 C. Hölzl, P. Kibies, S. Imoto, R. Frach, S. Suladze, R. Winter, D. Marx, D. Horinek, S. M. Kast, *J. Chem. Phys.* 144, 144104 (2016)

280 M. D Hanwell, D. E Curtis, D. C Lonie, T. Vandermeersch, E. Zurek, G. R Hutchison, *J. Cheminformatics*, 4, 17 (2012)

281 C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* 97, 10269 (1993)

282 P. Cieplak, W. D. Cornell, C. Bayly, P. A. Kollman, *J. Comput. Chem.* 16, 1357 (1995)

283 W. D. Cornell, P. Cieplak, C. I. Bayly, P. A. Kollmann, *J. Am. Chem. Soc.* 115, 9620 (1993)

284 J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graph. Model*. 25, 247260 (2006)

285 M. T. Geballe, J. P. Guthrie, *J. Comput.-Aided Mol. Des*. 26, 489 (2012)

286 D. L. Mobley, K. L. Wymer, N. M. Lim, J. P. Guthrie, *J. Comput.-Aided Mol. Des*. 28, 135 (2014)

287 C. C. Bannan, K. H. Burley, M. Chiu, M. R. Shirts, M. K. Gilson, D. L. Mobley, *J. Comput.-Aided Mol. Des*. 30, 927 (2016)

288 M. Işık, D. Levorse, A. S. Rustenburg, I. E. Ndukwe, H. Wang, X. Wang, M. Reibarkh, G. E. Martin, A. A. Makarov, D. L. Mobley, T. Rhodes, J. D. Chodera, *J. Comput.-Aided Mol. Des*. 32, 1117 (2018)

289 M. Işık, D. Levorse, D. L. Mobley, T. Rhodes, J. D. Chodera, *J. Comput.-Aided Mol. Des*. 34, 405 (2020)

290 https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubs.acs.org%2Farticlesonrequest%2FAOR-C98NW2AUQDMRQABN7AHQ&amp;data=02%7C01%7CRichardsN14%40cardiff.ac.uk%7C2527962c75ed4c2c46c208d7cccf7869%7Cbdb74b3095684856bdbf06759778fcbc%7C1%7C0%

7C637203063758411386&amp;sdata=F%2Biz-baojJvP12Usa%2BUuH8T4fR4ByIW2Y0ywKtK9kJRE%3D&amp;reserved=0

291 W. F. Reynolds, C. W. Tzeng, *Can. J. Bi-chem*. 55, 576 (1977)

292 R. E. Wasylishen, G. Tomlinson, *Can. J. Biochem*. 55, 579 (1977)

293 P. I. Nagy, G. J. Durant, W. P. Hoss, D. A. Smith, *J. Am. Chem. Soc*. 116, 4898 (1994)

294 F. Forti, C. N. Cavasotto, M. Orozco, X. Barril, F. J. Luque, *J. Chem. Theory Comput*., 8, 5, 1808 (2012)

295 M. Witanowski, L. Stefaniak, S. Szymanski, H. Junuszewski, *J. Magn. Reson*. 28, 217 (1977)

296 W. M. Litchman, M. Alei, A. E. Florin, *J. Am. Chem. Soc.* 91, 6574 (1969)

297 S. Maste, "Hybride implizite und explizite Solvatationsansätze zur Berechnung von NMR-Parametern", *Master's Thesis*, Technische Universität Dortmund, 2020

298 M. Alei, A. E. Florin, W. M. Litchman, J. F. O'Brien, *J. Phys. Chem*. 75, 7 (1971)

299 D. S. Wishart, C. G. Bigama, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfieldd, J. L. Marklet, B. D. Sykesa, *J. Biomol. NMR*, 6, 135 (1995)

300 G. Cao, W. Zheng, *Acta Phys.-Chim. Sin.* 29, 10, 2135 (2013)

301 E. D. Raczyńska, M. Makowski, K. Zientara-Rytter, K. Kolczyńska, T. M. Stępniewski, M. Hallmann, *J. Phys. Chem. A* 117, 1548 (2013)

302 A. O. Alyoubi, R. H. Hilal, *Biophys. Chem*. 55, 73 231 (1995)

303 J. R. Štocek, M. Dracínský, *Biomolecules* 10, 170 (2020)

304 J. Sepioł, Z. Kazimierczuk, D. Z. Shugar, *Z. Naturforsch., C: J. Biosci*. 31, 361 (1976)

305 K. N. Rogstad, Y. H. Jang, L. C. Sowers, W. A. Goddard, *Chem. Res. Toxicol*. 16, 1455 (2003)

306 T. A. Martinot, S. A. Benner, *J. Org. Chem*. 69, 3972 (2004)

307 D. Jiang, F. Seela, *J. Am. Chem. Soc*. 132, 4016 (2010)

308 D. Loakes, P. Holliger, *Chem. Comm*. 4619 (2009)

309 M. L. Malone, B. M. Paegel, *ACS Comb. Sci.* 18, 182 (2016)

310 M. Potowski, F. Losch, E. Wünnemann, J. K. Dahmen, S. Chines, A. Brunschweiger, *Chem. Sci*. 10, 10481 (2019)

311 A. S. Ratnayake, M. E. Flanagan, T. L. Foley, J. D. Smith, J. G. Johnson, J. Bellenger, J. I. Montgomery, B. M. Paegel, *ACS Comb. Sci.* 21, 650 (2019)

312 M. Pyrka, M. Maciejczyk, *Chem. Phys. Lett*. 627 30(2015)

313 C. Alhambra, J. Luque, J. Estelrich, M. Orozco, *J. Org. Chem*. 60, 969 (1995)

314 F. Zamora, M. Kunsman, M. Sabat, B. Lippert, *Inorg. Chem*. 36, 8, 1583 (1997)

315 J. Šponer, J. E. Šponer, L. Gorb, J. Leszczynski, B. Lippert, *J. Phys. Chem. A*, 103, 51, 11406 (1999)

316 B. Lippert, D. Gupta, *Dalton Trans*. 4619 (2009)

317 S. P. Samijlenko, Y. P. Yurenko, A. V. Stepanyugin, D. M. Hovorun, *J. Phys. Chem. B*, 114,1454 (2010)

318 V. H. Harris, C. L. Smith, W. J. Cummins, A. L. Hamilton, H. Adams, M. Dickman, D. P. Hornby, D. M. Williams, *J. Mol. Biol*. 326, 1389 (2003)

319 F. Seela, R. Kröschel, *Nucleic Acids Res*. 31, 24, 7150 (2003)

320 A. R. Morgan, *Trends Biochem. Sci*. 18, 1993

321 P. Khuu, P. S. Ho, *Biochemistry* 48,7824 (2009)

322 W. Wang, H. W. Hellinga, L. S. Beese, *Proc. Natl. Acad. Sci. U.S.A* 108, 43 17644 (2011)

323 Y. P. Wong, *J. Am. Chem. Soc*. 95,11, 1973

324 I. J. Kimsey, E. S. Szymanski, W. J. Zahurancik, A. Shakya, V. Xue, C.-C. Chu, B. Sathyamoorthy, Z. Suo. H. M. al-Hashimi, *Nature*, 554, 195 (2018)

325 E. Westhof, *FEBS Lett.* 558, 2464 (2014)

326 E. Pluharova, P. Jungwirth, S. E. Bradforth, P. Slavícek, *J. Phys. Chem. B*, 115, 1294 (2011)

327 L. E. Kapinos, B. P. Operschall, E. Larsen, H. Sigel, *Chem. Eur. J.* 17, 29, 8156 (2011)

328 W. C. Johnson, P. M. Vipond, J. C. Girod, *Biopolymers*, 10, 923 (1971)

329 R. Purrello, M. Molina, Y. Wang, G. Smulevich, J. Fossella, J. R. Fresco, T. G.Spiro, J. Am. Chem. Soc. 115 760 (1993)

330 C. M. Widdifield, R. W. Schurko, Concepts Magn. Reson., Part A, 34, 2, 91 (2009)

331 A. Barszczewicz, M. Jaszufiski, T. Helgaker, K. Ruud, *Chem. Phys. Lett*. 250, 1 (1996)

Appendix

Table 48: Molecules of the test-set, a subset of the Chamberlin dataset. The molecule name, minimum and maximum temperatures (in K) at which experimental free energies of solvation are available as well as the corresponding values (in kcal mol^{-1}) are given. Additionally, the linear regression parameters slope m and intercept b (in kcal mol^{-1}) are shown.

Molecule	min. temp.	min. $\Delta_{\text{solv}}G^0$	max. temp.	max. $\Delta_{\text{solv}}G^0$	b	m
1-ethoxybutane	273.15	-2.70	363.85	-0.13	-10.44	0.028
2-methylpentane	273.15	1.55	372.25	4.24	-5.86	0.027
2-propanol	273.00	-5.55	373.15	-3.03	-12.42	0.025
3-nitrophenol	273.15	-10.24	373.15	-8.50	-14.99	0.017
benzene	275.00	-1.45	373.14	0.07	-5.71	0.015
benzonitrile	273.15	-4.62	368.65	-3.06	-9.08	0.016
chloromethane	273.15	-1.02	373.15	0.31	-4.65	0.013
cyclopropane	293.15	0.69	318.15	1.18	-5.06	0.020
dimethylsulfide	273.15	-2.04	373.15	-0.49	-6.27	0.016
ethene	273.00	0.87	346.05	1.98	-3.28	0.015
ethylbenzene	273.15	-1.58	373.14	0.60	-7.54	0.022
ethylbenzoate	273.15	-4.42	363.45	-2.05	-11.59	0.026
fluoromethane	273.15	-0.60	373.15	0.50	-3.60	0.011
hexane	273.15	1.43	372.25	4.20	-6.20	0.028
hydrogensulfide	273.15	-0.85	373.15	-0.06	-3.01	0.008
methane	285.05	1.82	348.35	2.75	-2.37	0.015
methylamine	273.15	-5.11	373.15	-3.20	-10.33	0.019
morpholine	273.15	-7.99	373.15	-5.09	-15.91	0.029
octanoic_acid	273.15	-6.98	373.15	-3.87	-15.47	0.031
phenol	277.15	-6.79	373.15	-4.95	-12.10	0.019
piperidine	273.15	-6.07	373.15	-3.12	-14.13	0.030
propanethiol	273.15	-1.56	373.15	-0.06	-5.66	0.015
quinoline	274.15	-4.29	373.15	-2.50	-9.25	0.018
tetrafluoroethylene	273.15	1.30	373.15	3.09	-3.59	0.018
tetrahydrofuran	293.15	-3.58	343.15	-2.72	-8.62	0.017
trichloroethylene	274.95	-1.25	367.65	1.11	-8.25	0.025
urea	273.15	-14.54	373.15	-12.13	-21.12	0.024

Table 49: Results of the TI-MD calculations for the SAMPL2 dataset. The first column gives the molecular structures (given in the structures subfolder of SI part 02), followed by the free energy of solvation (Δ_{solv}G) and the corresponding error, as well as the separate non-polar (VdW) and electrostatic (Elec) components with their respective errors.

Molecule	$\Delta_{\text{solv}}G$	Error	$\Delta_{\text{solv}}G$(VdW)	Error	$\Delta_{\text{solv}}G$(Elec)	Error
10B_4	-9.24	0.19	0.82	0.13	-10.06	0.05
10D_4	-9.28	0.20	0.90	0.14	-10.18	0.06
12C_1	-9.27	0.22	1.21	0.17	-10.48	0.05
13D_1	-11.53	0.26	1.27	0.21	-12.80	0.05
14D_2	-5.93	0.18	1.39	0.13	-7.32	0.04
15C_1	-11.70	0.23	1.40	0.18	-13.09	0.05
1A_3	-5.26	0.17	1.10	0.12	-6.37	0.05
2B_1	-10.15	0.20	0.99	0.16	-11.14	0.04
4A_1	-11.09	0.23	1.05	0.17	-12.14	0.05
5A_2	-6.76	0.21	0.37	0.14	-7.13	0.07
6A_2	-8.65	0.21	0.48	0.16	-9.13	0.05
7B_cis_1	-4.04	0.16	2.14	0.12	-6.18	0.04
8A_1	-7.28	0.17	1.17	0.14	-8.44	0.03
10B_5	-12.63	0.18	0.96	0.13	-13.59	0.06
11C_1	-9.31	0.18	1.12	0.14	-10.43	0.04
12D_1	-9.97	0.22	1.52	0.16	-11.49	0.05
13D_2	-7.50	0.23	1.65	0.18	-9.15	0.05
15A_1	-7.12	0.22	1.47	0.18	-8.59	0.04
16A_1	-6.29	0.17	1.00	0.14	-7.29	0.03
1B_1	-9.80	0.16	0.94	0.12	-10.75	0.04
3A_1	-9.93	0.20	1.10	0.16	-11.03	0.05
4A_3	-6.16	0.21	1.02	0.16	-7.18	0.05
5B_1	-12.69	0.18	0.28	0.13	-12.97	0.05
6B_1	-9.43	0.22	0.28	0.18	-9.71	0.04
7B_cis_3	-8.02	0.17	2.07	0.13	-10.09	0.05
8B_1	-4.64	0.18	1.38	0.15	-6.02	0.04
10C_1	-10.47	0.18	0.89	0.14	-11.37	0.05
11D_1	-10.82	0.21	1.24	0.16	-12.06	0.05
12D_2	-7.03	0.22	1.58	0.17	-8.61	0.05
14C_1	-7.15	0.18	1.12	0.14	-8.27	0.04
15B_2	-6.70	0.22	2.11	0.18	-8.80	0.04
16C_1	-9.41	0.17	0.94	0.13	-10.35	0.04
2A_1	-11.18	0.21	1.17	0.16	-12.35	0.05
3A_4	-6.52	0.20	0.91	0.15	-7.43	0.05
4B_1	-12.12	0.21	0.49	0.16	-12.61	0.05
5C_1	-16.42	0.18	0.31	0.13	-16.73	0.05
6Z_1	-20.32	0.22	0.03	0.16	-20.35	0.05
7B_trans_1	-5.80	0.16	2.22	0.12	-8.02	0.04
8B_3	-8.79	0.18	1.16	0.14	-9.95	0.04
10D_1	-12.87	0.19	0.94	0.13	-13.81	0.05
11D_4	-7.03	0.21	1.45	0.16	-8.48	0.05
13C_1	-8.79	0.23	1.56	0.19	-10.35	0.04
14D_1	-8.12	0.22	1.48	0.17	-9.60	0.05
15B_4	-9.64	0.25	1.33	0.20	-10.97	0.05
1A_1	-9.90	0.18	1.28	0.13	-11.18	0.05
2A_2	-6.49	0.21	0.97	0.16	-7.46	0.05

Molecule	$\Delta_{\text{solv}}G$	Error	$\Delta_{\text{solv}}G$(VdW)	Error	$\Delta_{\text{solv}}G$(Elec)	Error
3B_1	-10.00	0.22	0.65	0.18	-10.65	0.04
5A_1	-11.15	0.19	0.72	0.13	-11.87	0.06
6A_1	-10.40	0.20	0.61	0.16	-11.01	0.04
7A_g1_1	-4.10	0.15	2.08	0.12	-6.18	0.03
7B_trans_2	-5.33	0.17	2.14	0.13	-7.47	0.04

Table 50: Experimental histamine NMR parameter and the assignment to the respective atoms. The ^{1}H and ^{13}C shifts refer to TMS in deuterated methylene chloride; for the ^{15}N shifts, the spectrometer calibration to nitromethane was used.[277]

Sample	**Parameter**	**α-C**	**β-C**	**C-2**	**C-4**	**C-5**
pH 7.9	δ_C	39.921	25.114	116.886	136.945	133.773
pH 12.5	δ_C	40.98	29.72	117.841	136.586	135.711
		α-CH₂	**β-CH₂**	**H-2**	**H-4**	**-**
pH 7.9	δ_H	3.244	2.934	7.025	7.736	-
pH 12.5	δ_H	2.807	2.653	6.866	7.624	-
		α-N	**τ-N**	**π-N**	**-**	**-**
pH 7.9	δ_N	-350.4	-205.93	-151.7	-	-
pH 12.5	δ_N	-358.33	-191.56	-162.35	-	-
		$^{3}J_{\alpha\text{-CH}_2/\beta\text{-CH}_2}$	$^{3}J_{\beta\text{-CH}_2/\alpha\text{-CH}_2}$	$^{4}J_{\text{H-2/H-4}}$	$^{4}J_{\text{H-4/H-2}}$	**-**
pH 7.9	J_{HH}	7.025	6.96	0.9	1.05	-
pH 12.5	J_{HH}	6.865	6.835	0.7	0.94	-
		$^{1}J_{\alpha\text{-C}/\alpha\text{-H}_2}$	$^{1}J_{\beta\text{-C}/\beta\text{-H}_2}$	$^{1}J_{\text{C-2/H-2}}$	$^{1}J_{\text{C-4/H-4}}$	**-**
pH 7.9	J_{CH}	144.9	129.4	191.2	209.4	-
pH 12.5	J_{CH}	136.4	127	189.1	207.8	-