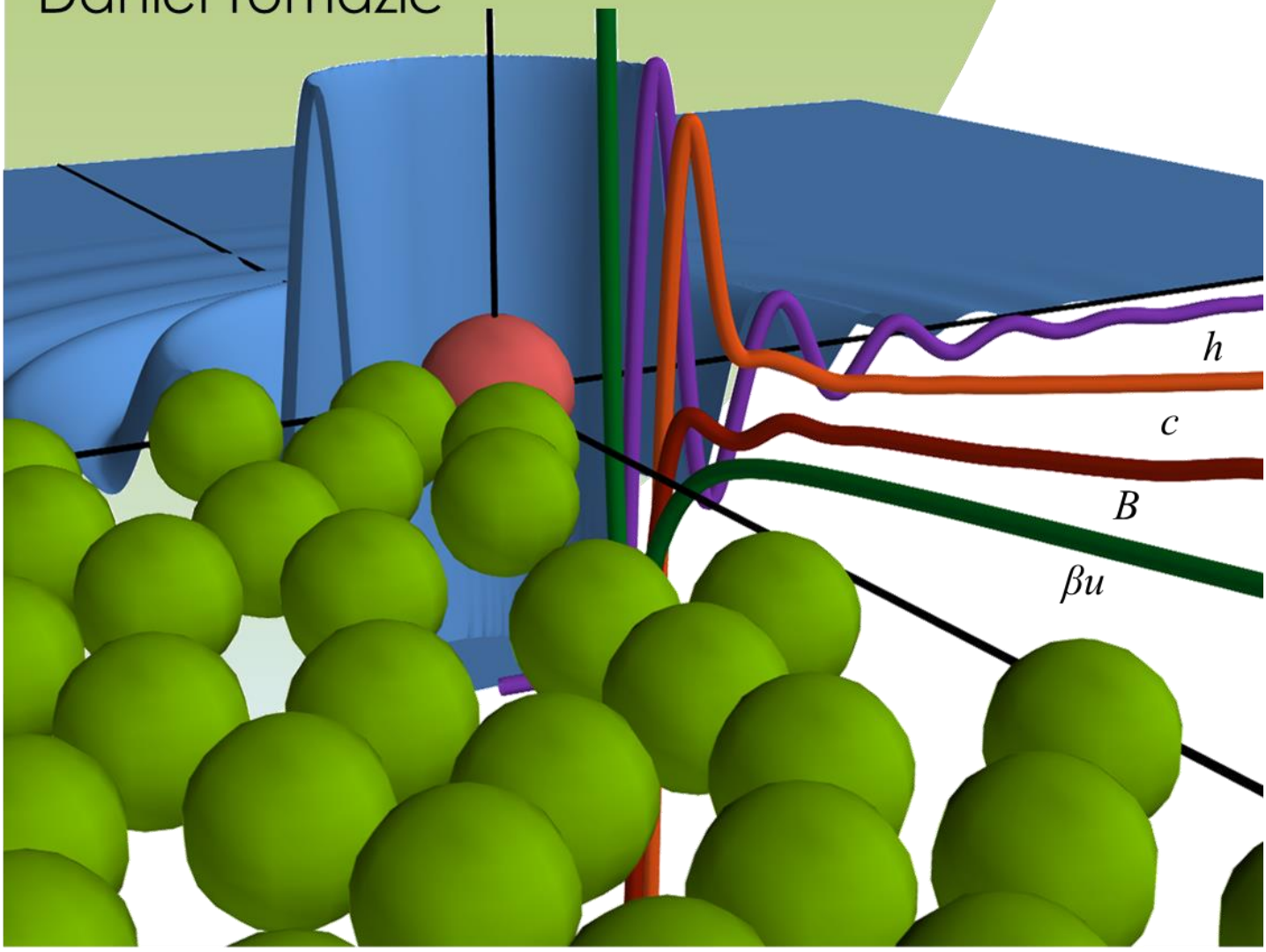
Fakultät für Chemie und Chemische Biologie
Technische Universität Dortmund

Optimizing free energy functionals in integral equation theories

Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) vorgelegt im Februar 2016 von

Daniel Tomazic



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

Danksagung

Bedanken möchte ich mich zuallererst bei Herrn **Professor Dr. Stefan M. Kast**. In seinem Arbeitskreis hatte ich die Gelegenheit sehr eigenständig zu arbeiten. Stefan war jederzeit mit gutem Rat zur Stelle, wenn Probleme auftauchten.

Für das Anfertigen des Zweitgutachtens bedanke ich mich herzlich bei Herrn **Professor Dr. Roland Winter**.

Mein Dank gilt auch **Patrick Kibies, Martin Urban, Jochen Heil, Franziska Hoffgaard, Nicolas Tielker, Roland Frach, Florian Mrugalla, Leonhard Henkes** und den anderen Mitgliedern des Arbeitskreises. Ich habe mich bei euch immer sehr wohl gefühlt, was mir den morgendlichen Weg zur Arbeit nochmals erleichtert hat. Sehr genossen habe ich die Zeit in den Mittagpausen bei den Kaffee-Kollegen. Hier konnten in entspannter Atomsphäre berufliche und andere Dinge zwanglos besprochen werden. Besonders bedanken möchte ich mich an dieser Stelle bei Frau **Anneliese Ahlke**, die mir und dem Rest der Gruppe viel Bürokratie abgenommen hat.

Ein ganz besonderer Dank geht an all jene, die mir außerhalb der Universität so sehr ans Herz gewachsen sind und bei denen ich immer je nach Bedarf Freude, Halt und Trost gefunden habe. Das ist zuallererst meine Freundin **Natascha** (die ich während meiner Promotionszeit erst kennen gelernt habe), meine alten Freunde aus der Schulzeit **Phillip, Moritz, Lars** und **Dennis** sowie meine ehemaligen Kommilitonen **Arthur, Micha, Julia** und **Sebastian**. Dazu kommen noch meine Freunde aus den Football Teams der Recklinghausen Chargers und der Gelsenkirchen Devils.

Den tiefsten Dank empfinde gegenüber meinen Eltern **Monika** und **Jörg**. Beide haben seit ich denken kann weder Kosten noch Mühen gescheut, mir meinen Lebensweg zu ermöglichen. Jörg hat mich bei der vorliegenden Arbeit sehr unterstützt, indem er sich mit mir durch den Text geackert hat. Ich froh, dass ich euch habe!

Daniel

Diese Arbeit entstand zwischen dem 01.09.2011 und 31.12.2015 unter der Leitung von Herrn Prof. Dr. Stefan M. Kast. Sie wurde von M. Sc. Daniel Tomazic in Fakultät für Chemie und chemische Biologie der Technischen Universität Dortmund im Arbeitskries für theoretische physikalische Chemie angefertigt.

Die hier vorgestellten Ergebnisse stammen vom Autor. An den folgenden Stellen wurde auf die Arbeit anderer zurückgegriffen.

- Der pK_a Datensatz wurde von Jochen Heil aus Litheraturdaten zusammengestellt er hat er die Geometrieoptimierungen dieser Moleküle durchgeführt.
- Das Programm zum konformationellen *sampling* wurde von Patrick Kibies geschrieben und von mir um die Schnittstelle zu RDKit erweitert.
- Die Definition des P4G Modells stammt von Nikolas Tielker
- Das Programm zum Lösen der RISM Gleichungen wurde von Thomas Kloss und Jochen Heil basierend auf einer Vorlage von Stefan Kast geschrieben. Mein Beitrag zu diesem Programm beschränkt sich auf die Implementierung der HNC-BF0 *closure*.
- Die EC-RISM Rechnungen wurden mit einem Programm von Jochen Heil, Thomas Kloss und Patrick Kibies durchgeführt.
- Die Suszeptibilitätsfunktionen von Wasser wurden von Stefan Kast und Thomas Kloss (HNC) sowie von Stefan Kast und Patrick Kibies (MD) angefertigt.

Teile der hier vorgestellten Ergebnisse sind bereits Teil der Veröffentlichung:

Tomazic, D.; Hoffgaard, F.; Kast, S. M. *Chemical Physics Letters* 2014, *591*, 237-242.

Das Bild der Wasseroberfläche, das in einigen Schemata zu sehen ist stammt von Tobias Pfau (www.tobiaspfau.de) und wurde mit seiner Zustimmung verwendet.

Eidesstattliche Versicherung (Affidavit)

Tomazic, Daniel
Name, Vorname
(Surname, first name)

108828
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:

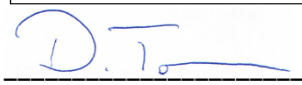
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of 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.

Dortmund, 03.02.2016
Ort, Datum
(Place, date)


Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Optimizing free energy functionals in integral equation theories

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.

Dortmund, 03.02.2016
Ort, Datum
(Place, date)

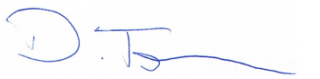

Unterschrift
(Signature)

Table of contents

1 Zusammenfassung

Das *reference interaction site model* RISM[1,2,3,4] gewinnt für die Berechnung von Solvenzstruktur und -thermodynamik zunehmend an Bedeutung. Bei RISM handelt es sich um ein statistisches Lösungsmittelmodell, das zu den impliziten Lösungsmittelmodellen gehört. Dieses basiert auf der Integralgleichungstheorie. Eine seiner zentralen Größen ist die *bridge function*, welche die Qualität der Berechnungen deutlich beeinflusst. Ungeachtet ihrer Bedeutsamkeit ist die *bridge function* in der Regel nicht bekannt und wird lediglich angenähert. Diese Arbeit befasst sich mit der Erhöhung der Genauigkeit mit Hilfe von RISM berechneter Lösungsmittelthermodynamiken. Der Einfluss der *bridge function* auf das chemische Exzesspotential wird untersucht. Im Anschluss werden die Ergebnisse dazu verwendet, eine vielfältig einsetzbare Korrektur für mittels RISM berechnete freie Enthalpien zu erstellen. Die sich stellenden Fragen sind:

- Was ist der Effekt der *bridge function* auf die freie Solvatationsenthalpie?
- Wie lässt sich dieser Effekt präzise quantifizieren?
- Kann dieses auf Systeme mit unbekannten/genäherten *bridge functions* übertragen werden?

Diese Arbeit über die Berechnung des Beitrags der *bridge function* zum chemischen Exzesspotential ist in zwei Teile gegliedert. Der erste Teil befasst sich mit einfachen Modellfluiden. Die direkten und indirekten Einflüsse der *bridge function* werden anhand des Lennard-Jones (LJ) und des repulsiven Weeks-Chandler-Anderson (WCA) Fluids untersucht. Anhand dieser simplen Modelle werden verschiedene geschlossene, *bridge*-abhängige Funktionale und deren numerische Stabilität diskutiert. Da die *bridge function* des WCA-rep Fluids bis dahin unbekannt war, wurde sie mit Moleküldynamik-Simulationen bestimmt. Im zweiten Teil werden die Ergebnisse aus dem ersten Teil auf wässrige Lösungen übertragen. Hierzu wird eine semi-empirische Korrektur der freien Hydratationsenthalpien für das *embedded cluster reference interaction site model* (EC-RISM[5]) parametrisiert.

In dieser Arbeit werden drei Themenbereiche behandelt:

Berechnung von *bridge functions*: Mittels Molekulardynamik-Simulationen wurde die *bridge funiiton* des WCA-rep Fluids bestimmt. Das durch das Lösungsmittel bedingte Potenzial der mittleren Kräfte, das auf ein Dimer wirkt, wurde verwendet um die *bridge function* bei kleinen Abständen zu ermitteln. Die *bridge function* des WCA-rep Fluids wurde mit der des LJ Fluids verglichen. Zwei verschiedene Methoden zur Invasion der Orstein-Zernike Ge-

lichung wurden gegenübergestellt. Die *bridge function* wurde bei unterschiedlichen Dichten und Temperaturen berechnet, die jeweils flüssigen Zuständen des LJ Fluids entsprechen.

Berechnung chemischer Exzesspotentiale mit *bridge functions:* Mithilfe der *bridge functions* wurde geschlossene von der *bridge function*-abhängige Funktionale evaluiert um das chemische Exzesspotential zu bestimmen. Das chemische Excesspotential beim Übergang des WCA-rep Fluid zum LJ Fluid konnte mit hoher Genauigkeit berechnet werden. Die HNC *closure* erwies sich als eine gute Näherung als sie um ein *bridge*-abhängiges Funktional erweitert wurde. Es wurden zwei verschiedene geschlossene Ausdrücke untersucht. Obwohl beide Korrekturen nicht erfolgreich waren, ähnelten sich die Dichten der chemischen Exzesspotentiale. Der Effekt der *bridge function* auf das chemische Exzesspotential wurde anhand des Zero-Separation-Theorems harter Kugeln[6,7,8,9] untersucht.

Semi-empirische Korrekturen: Nach der Multiplikation mit einem semi-empirischen Faktor entspricht das partielle Molvolumen näherungsweise dem *bridge*-abhängigen Funktional. Dies gilt besonders für neutrale Moleküle und bei Ionen muss ein weiterer Ladungsabhängiger Faktor hinzugefügt werden. Mit dieser Korrektur kann RISM mit anderen modern impliziten Lösungsmittelmodellen vergleichbare Ergebnisse erzielen. Die Rechnungen mit Ionen sind sogar genauer. Um eine optimale Korrelation mit experimentellen Daten zu erhalten wurden weitere Korrekturfaktoren getestet. Die Vorhersagekraft der Modelle wurde anhand mehrerer Trainings- und Testdatensäte validiert. Die Korrektur wurde für verschiedene quantenmechanische Theorieniveaus und Basissätze sowie verschiedene *closures* und Suszeptibilitätsfunktionen parametrisiert. Ähnlich den SM*x*[10,11,12] Modellen können auch bei der partiellen Molvolumenskorrektur im Vakuum optimierte Molekülgoemotrien verwendet werden ohne, dass diese weiterer Optimierungen bedürfen. Einer Untersuchung der chemischen Zusammensetzung ergab, dass die freie Energiekorrektur am besten mit Molekülen funktioniert, die Sauerstoff und Chlor enthalten. Die Korrelation mit den experimentellen Daten verschlechtert sich wenn das Molekül Stickstoff-, Fluor-, Brom- oder Phosphoratome enthält.

Die Ergebnisse zeigen, dass RISM zur genauen Berechnung chemischer Exzesspotentiale und freier Hydratationsenthalpien geeignet ist. Die HNC[13,14] *closure* und PSE-*k*[15] *closures* profitieren vom Hinzufügen eines bridge-abhängigen Terms zu ihren chemischen Exzesspotentialfunktionalen. Da eine Verwendung der *bridge function* häufig zur Divergenz von RISM-Rechnungen führt, ist die gezeigte nachträgliche Korrektur der freien Enthalpie vorteilhaft.

1 Abstract

For the calculation of solvation thermodynamics and structures the reference interaction site model RISM[1,2,3,4] is a theory with raising importance. RISM is a statistical solvent model that belongs to implicit solvation models. It is based on integral equation theory. A central property in RISM is the bridge function which heavily influences the quality of the calculations. Regardless of its importance the bridge function is usually unknown and approximations are used instead. This work is about improvement of the accuracy of solvation thermodynamics properties calculated with RISM. The influence of the bridge function on the excess chemical potential is studied and findings are used to build a widely usable correction for RISM calculated free energies. Hence the questions of this work are:

- What is the influence of the bridge function on the free energy of solvation?
- How can this influence be quantified accurately?
- Can this be transferred to systems with unknown/approximate bridge functions?

This work about the calculation of bridge function contributions to the excess chemical potential is divided into two parts. Part one is about simple model fluids. Here the bridge functions of the Lennard-Jones (LJ) fluid and the repulsive Weeks-Chandler-Anderson (WCA-rep) fluid were used to study the direct and indirect bridge function contributions. At the example of these simple model fluids certain bridge dependent closed form expressions for the excess chemical potential and their numerical stability are discussed. Since the bridge function of the WCA-rep was unknown so far, it was calculated from molecular dynamics (MD) simulations. The second part is about a generalization of the results from part one to aqueous solutions. Here a semi-empirical correction of hydration free energies is parametrized for the embedded cluster reference interaction site model (EC-RISM[5]).

In this thesis three main subjects have been discussed:

Calculation of Bridge functions: The bridge function of the WCA-rep fluid was calculated using molecular dynamics simulations. The solvent induced potential of mean forces on a constrained dimer was used to extract the bridge near zero separation. The bridge function of the WCA-rep fluid is compared to the LJ fluid. The Ornstein–Zernike equation was inverted in different ways (directly in reciprocal space or iteratively in real space) to find the most robust methodology. The bridge dependent functional is based on a parametric representation of the bridge function which is influenced by the thermodynamic state. So the calculations were

performed at different densities and temperatures corresponding to liquid states of the LJ fluid.

Calculation of excess chemical potentials with bridge functions: The information about the bridge function was used to calculate excess chemical potentials with closed form bridge dependent expressions for the excess chemical potential. The transition free energies from the WCA-rep fluid to the LJ fluid were calculated with high accuracy. In combination a bridge dependent functional the HNC closure proved to be a good approximation for the exact closure. Two different closed form expressions were tested to calculate exact chemical potentials of the LJ fluid. Though both corrections were unsuccessful, the excess chemical potential density was similar. With the zero separation theorem[6,7,8,9] of hard sphere fluids the effect of the bridge function on the excess chemical potential was studied.

Semi-empirical corrections: The partial molar volume can be used to approximate the bridge dependent functional by using a semi-empirical scaling factor. This works well for neutral molecules. For ions an additional charge dependent term is needed. This correction is competitive with other state of the art implicit solvent models especially in the treatment of ions. It was tested which additional correction terms can be used to further improve the agreement of the calculations with experimental data. The predictive power was confirmed by multiple reparametrizations with training and test compounds. The partial molar volume correction was benchmarked with various combinations of quantum mechanical levels of theory and basis sets. Also different closures and susceptibility functions within the reference interaction site model were tested. Similar to the SM*x*[10,11,12] models vacuum optimized structures can be used for the calculations with great success without the need of further geometry optimizations. A classification of the molecules by atom types shows that the correction works best for molecules containing oxygen and chlorine. The correlation with the experimental solvation free energies is slightly worse when nitrogen, fluorine, bromine or phosphorous are part of the molecule.

The results from the presented studies show that the reference interaction site model can be used for the accurate calculation of excess chemical potentials and solvation free energies. The HNC[13,14] closure and the PSE-*k*[15] closures greatly benefit from the addition of a bridge dependent term to their respective excess chemical potential functional. An addition of the bridge function to the RISM iteration often leads to a divergence of the calculation which may be a consequence of a validation of a mathematical constraint called LambertW[16] condition. Hence the subsequent correction of converged calculations is beneficial. The semi-empirical correction presented here is well suited for this purpose.

2 Introduction

A major part of chemistry takes place in solution. Together with experimental measures a theoretical description of solvation phenomena is an approach to gain insight into solvation related subjects. In a theoretical framework solvation can be modeled in two different ways. In the explicit solvation models all solvent molecules are treated individually. These models can be very accurate but they are also time-consuming. An alternative are the implicit solvent models which describe the solvent by macroscopic properties. These models provide the results faster but include more approximations. Beside the popular continuum solvation models, the statistical solvation models like the reference interaction site model RISM belong to the implicate solvation models. RISM describes the solvent by means of local density fluctuations.

The major goal of this work is to increase the accuracy of thermodynamic predictions made with the reference interaction site model. This will be done by means of two different approaches. The first part uses the best possible physical description of simple model fluids to correct excess chemical potentials. The second part is about the description of aqueous systems and uses a semi-empirical correction based on the results of part one.

First, the physical basis of the bridge function and bridge function dependent corrections of the excess chemical potential are described. This function is a central quantity in the reference interaction site model. It is hard to measure and its influence is not completely understood yet. The bridge functions of two model fluids are compared and bridge function properties are discussed from a thermodynamic perspective. Tasks in this part are to

- implement the formalism for the calculation of bridge function contributions,
- identify the effects of the bridge function on the excess chemical potential,
- generalize the findings from the models for the usage of real solvents and solutes.

In the second part the findings of the model fluid studies are transferred to aqueous solutions. In this part the focus is shifted to the applicability of the reference interaction side model. A method is proposed that corrects RISM calculated solvation free energies of small molecules without increasing the computational burden. Experimental solvation free energies are used to parameterize this semi-empirical correction. Tasks of this part are to

- correct RISM calculated solvation free energies,
- estimate the expected error,
- build a reproducible workflow featuring a small expected error.

3 Theoretical Background

A large amount of chemical reactions and processes takes place in solution. In this chapter some basic concepts of solvation and theoretical approaches towards them will be presented.

3.1 Free energy of solvation

Chemical reactions are influenced by the environment they take place in. The presence of surrounding atoms and molecules and their physical properties influence the path a reaction follows. Hence a proper description of the educts, products and the surrounding medium is necessary for the prediction of chemical processes like reactions, binding, enrichment, aggregation, pattern formation and many more.

Referring to chemical reactions, a central quantity within those predictions is the Gibbs free energy of reaction $\Delta_R G°$. It is defined as the difference between products and educts. In the gas phase this property only depends on the properties of educts and products of a reaction. In solutions also the solvent has to be taken into account. This is often done by construction of a thermodynamic cycle, dividing the reaction into simpler sub reactions (see figure 1).

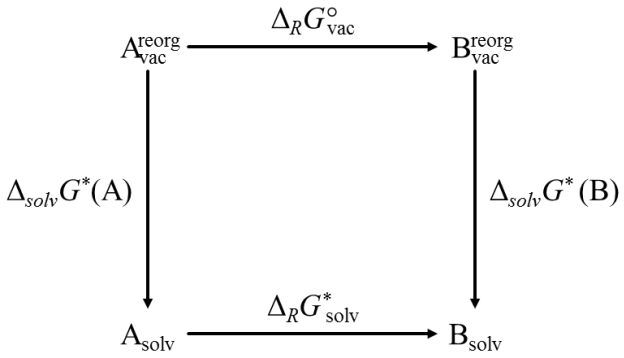


Figure 1: Thermodynamic cycle of the reaction A→B in the gas phase (top) and in solution (bottom).

A reaction in solution may be separated ($A_{solv} \rightarrow B_{solv}$) into three sub-reactions:

- reaction of the educts to the products ($A_{vac} \rightarrow B_{vac}$) in the gas phase,
- transfer of all educts into the gas phase ($A_{solv} \rightarrow A_{vac}$),
- transfer of products from the gas phase into solution ($B_{vac} \rightarrow B_{solv}$).

The free energy related to the transfer of a molecule from the gas phase into a solvent is called free energy of solvation. Assuming that the vibrational and rotational contributions are negli-

gible (they approximately cancel), the free energy of solvation of a molecule with a frozen geometry can be approximated by[5]

$$\Delta_{solv}G = \Delta_{solv}E^{\text{reorg}} + \Delta\mu\,. \tag{3.1}$$

Here $\Delta_{solv}E^{\text{reorg}}$ is the solvent introduced polerization energy of the solute due to an elecrionc reorgnisation. It is calculated from $\Delta_{solv}E^{\text{reorg}} = E_{\text{solv}} - E_{\text{vac}}$ wich are the quantum mechanical energies in solution and in the gas phase. The excess chemical potential (of a one component fluid) μ^{ex} is defined as

$$\mu^{\text{ex}} = \mu - \mu^{\text{id}}\,. \tag{3.2}$$

So it is the difference between the chemical potential of the system and the chemical potential of the corresponding ideal gas. The chemical potential is defined as

$$\mu = \left(\frac{\partial G}{\partial N}\right)_{p,T} = \left(\frac{\partial A}{\partial N}\right)_{V,T} = \ldots \tag{3.3}$$

with N being the number of particles in the system, V, p and T being the volume, pressure and the temperature respectively.

The excess chemical potential describes the change of the free energy when introducing an infinitesimally small change in the number of particles in the solvent in a frozen state. Since the smallest possible change to a system is adding (or removing) a single particle, the excess chemical potential defines the process of transferring a particle from an interaction free environment (gas phase) into the solvent. Since the excess chemical potential has to be calculated for all molecules in the reaction, the ability to predict it fast and accurately is desirable.

The free energy depends on the conditions under which the process is executed. Hence standard conditions were defined (for example by the IUPAC). In this work "°" denotes a quantity at 298.15 K and 1 atm for a process in the gas phase and 298.15 K and 1 mol/l for a process in solution, which is a common standard state in the American literate. The IUPAC convention uses 1 bar as the standard pressure. Another convention used is the Ben-Naim standard state[17], it will be denoted with a "#". It defines the concentration as 1 M in both phases and assumes ideality. This means a given free energy value corresponds to an ideal gas at a concentration of 1 mol/l dissolving as an ideal solution at a concentration of 1 mol/l. The solvent is treated as an ideal solution[17,18].

Due to this convention the ideal terms of the chemical potential cancel when inserting equation (3.2) in equation (3.1). Only the excess chemical potential remains

$$\Delta_{solv}G = \Delta_{solv}E^{\text{reorg}} + \mu^{\text{ex}}. \tag{3.4}$$

The transition free energy from one state to the other is given by the free energy difference of an ideal gas at 1 atm and 1 mol/l. It can be calculated from

$$\Delta G^{\circ\to\#} = -RT\ln(K) = RT\ln(\frac{c^{\circ}}{c^{\#}}). \tag{3.5}$$

The concentration coefficient is given by

$$\frac{c^{\circ}}{c^{\#}} = \frac{1\frac{\text{mol}}{\text{l}}}{\frac{1\,\text{atm}}{RT}} = \frac{1\frac{\text{mol}}{10^{-3}\,\text{m}^3}}{\frac{1.013310^5\ \text{J m}^{-3}}{8.314\,\text{J mol}^{-1}\text{K}^{-1}\ 298.15\,\text{K}}}. \tag{3.6}$$

Here R is the ideal gas constant. This results in

$$\Delta G^{\circ\to\#} = RT\ln(24.46). \tag{3.7}$$

It is possible calculate solvation free energies at both standard states form the energies and chemical potentials of single particles (see chapters 3.3 and 3.4 for information about calculations of molecular energies)

$$\begin{aligned}\Delta_{solv}G^{\#} &= \Delta_{solv}E^{\text{reorg}} + \mu^{\text{ex}}\\ \Delta_{solv}G^{\circ} &= \Delta_{solv}G^{\#} - \Delta G^{\circ\to\#}.\end{aligned} \tag{3.8}$$

From chapter 4 on all thermodynamic quantities will use the Ben-Naim standard and the "#" will be left out for the sake of readability.

When a molecule is flexible, there may be more than one conformation that contributes to the free energy. In this case the contributions of all conformations C can be averaged using partition functions to generate a conformational free energy[19]

$$\begin{aligned}G_{solv} &= -RT\ln\sum_{C}\exp[-G_{\text{C}}^{\text{sol}}/RT],\\ G_{vac} &= -RT\ln\sum_{C}\exp[-G_{\text{C}}^{\text{vac}})RT],\\ \Delta_{solv}G &= G_{solv} - G_{vac}.\end{aligned} \tag{3.9}$$

As described in the next chapters, the free energy of solvation can be calculated using various methods and theories. Most of these models include some kind of parametrization using experimental datasets. In this case the ensemble contributions are often implicitly included into the parametrization (see [20] and [21] for further discussions).

In a solvation process under the Ben-Naim standard conditions the excess chemical potential does not depend on the ensemble. Hence the change of the Helmholtz free energy $\Delta_{solv}A^{\#}$ equals the change of the free energy $\Delta_{solv}G^{\#}$.

3.2 Free energy of reactions

When looking at the thermodynamic cycle in figure 1 the excess chemical potential is used on the left hand site for a process in the direction from solvent to vacuum and on the right hand site for the other direction. This means that only the difference of the excess chemical potentials $\Delta\mu^{ex}(A\rightarrow B)=\mu^{ex}(A)-\mu^{ex}(B)$ and difference of electronic energies in the solvent $\Delta E_{solv}(A\rightarrow B) = E_{solv}(B) - E_{solv}(A)$ are of importance.

A direct calculation of $\Delta\mu^{ex}(A\rightarrow B)$ by estimation of the excess chemical potentials of both reactants will be systematically wrong in most cases. When A and B are similar compounds these errors may compensate when the difference is calculated. This means the calculation of free energies of reactions is often more accurate than the calculation of raw free energies of solvation.

To maximize the effect of error compensation in the prediction of free energies of reactions it is beneficial to focus on reactions that only introduce a small change to the observed molecule. Good examples for these types of reactions are proton transfer reactions. Therefore prediction of acidity constants (pK_a) or tautomer ratios are good benchmark cases for the calculation of free energies. The acidity constants of many substances have be measured so for. Therefore they are often used as an experimental reference. Since acidity constants are familiar to most chemists, they will be used as an example for the calculations of free energies and excess chemical potential and polarization energies in the chapter. Also at the end of this chapter some statistical measures will be given to evaluate the quality of calculated free energies.

An alternative to the measurement of acidity constants is the *in silico* calculation. The calculation is necessary when the measurement is complicated or even impossible with state of the art techniques. Therefore the calculation of acidity constants is an important task for computational chemistry. It can be used for theory benchmarks as well as for actual research. The thermodynamic cycle used for the calculation of an acidity constant is shown in figure 2.

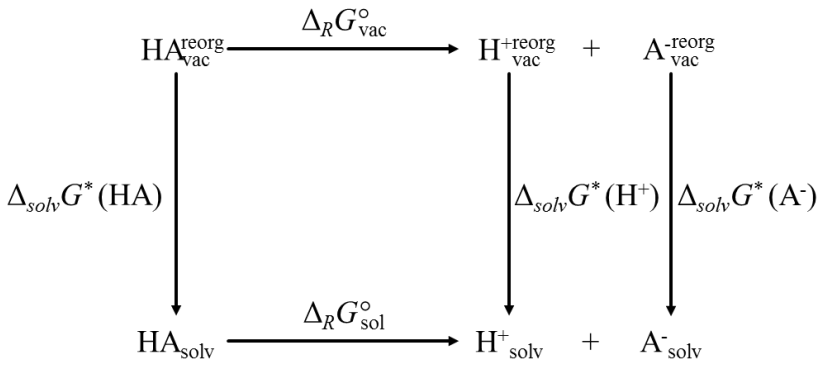


Figure 2: Thermodynamic cycle of the dissociation of the acid HA.

The free energy of the dissociation in gas phase $\Delta_R G^\circ_{vac}$ is usually approximated by the difference in the electronic energy of the acid HA and the corresponding base A^-. This energy difference can be calculated via quantum mechanics. The solvation free energies are calculated for all educts and products applying any kind of solvation model.

Thus the free energy of the reaction is given by

$$\begin{aligned}\Delta_R G = &-E_{\text{solv}}(\text{HA}) - \mu^{\text{ex}}(\text{HA}) \\ &+ E_{\text{solv}}(\text{A}^-) + \mu^{\text{ex}}(\text{A}^-) + E_{\text{solv}}(\text{H}^+) + \mu^{\text{ex}}(\text{H}^+)\end{aligned} \tag{3.10}$$

and all gas phase energy terms cancel. The excess chemical potential of the proton $\mu^{\text{ex}}(H^+)$ cannot be computed by the means of electronic structure calculation methods or measured directly. However, it can be extrapolated from a series of ion water cluster solvation free energies. Usually the experimental values of 265.9 kcal/mol by Tissandier *et al.*[22] for the solvation free energy of the proton and of 6.28 kcal/mol by Liptak and Schields[23] for the vacuum free energy of the proton are used.

When the free energy is known as seen in (3.10) the acidity constant is given by

$$\text{p}K_\text{a} = \frac{1}{RT\ln(10)}\Delta_R G^\circ . \tag{3.11}$$

When benchmarking theoretical calculations, a comparison of calculated and experimental free energies can be derived from equations (3.10), (3.7) and (3.11) by

$$\begin{aligned}\Delta_R G^\#_{\text{exp}} &= RT\ln(10)\text{p}K_\text{a} + \Delta G^{\circ\to\#} \\ \Delta_R G^\#_{\text{calc}} &= -E_{\text{solv}}(\text{HA}) - \mu^{\text{ex}}(\text{HA}) + E_{\text{solv}}(\text{A}^-) + \mu^{\text{ex}}(\text{A}^-) \\ &\quad + 265.9\,\text{kcal/mol} + 6.28\,\text{kcal/mol}.\end{aligned} \tag{3.12}$$

Other approaches towards the calculation of acidity constants are used that do not rely on experimental energies of the proton. In this case multiple acidity constants are used instead. As an example differences of acidity constants or a linear regression models can be used[19,24].

When only the difference of the acidity constants of two acids is needed (ΔpK_a or also called pK_a-Shift), the contribution of the proton cancels.

To calculate the acidity constant of a target acid with this method, a reference acid with a known acidity constant is needed. In this case ΔpK_a of target and reference acid can be calculated. Due to error compensation, this method works best when target and reference are similar.

When the pK_a-values of several reference compounds are known, a linear regression model can be applied. In this case a "raw" pK_a is calculated from

$$pK_a(\text{raw}) = \frac{-1}{RT\ln(10)}(-E_{solv}(\text{HA}) - \mu^{ex}(\text{HA}) + E_{solv}(\text{A}^-) + \mu^{ex}(\text{A}^-)). \quad (3.13)$$

Then the "raw" pK_a is used in a linear model with the experimental values[19,24]

$$pK_a(\text{exp}) \sim a\, pK_a(\text{raw}) + b. \quad (3.14)$$

Additionally the quality of the model can be measured by this approach. The ideal value of *a* is one and the ideal value of *b* is correlated with the excess chemical potential of the proton. Additional measures for the quality of a linear model are the root mean squared error RMSE and the coefficient of determination R^2. These measures will be used in chapter 7.

Another important process is the complexation of a solute. Complexation is important in many fields such as catalysis and biochemistry. For a physical or theoretical chemist, complexation of a solute by the solvent is a tool to study solvation phenomena and benchmarking of solvation methods which will be discussed in the next chapter.

The experimental value of the solvation free energy of the proton by Tissandier *et al.*[22] was calculated using the so called cluster pair approximation[22,25,26]. In this method the solvation free energy of ion solvent clusters are measured for different ions and different numbers of complexing molecules to extrapolate the solvation free energy of the bare proton.

As shown in figure 3, the complexation free energy of a molecule in the solvent by an explicit solvent molecule is expected to be zero. This cycle can be used to validate solvation state theories as discussed in chapter 7.3.3.2.

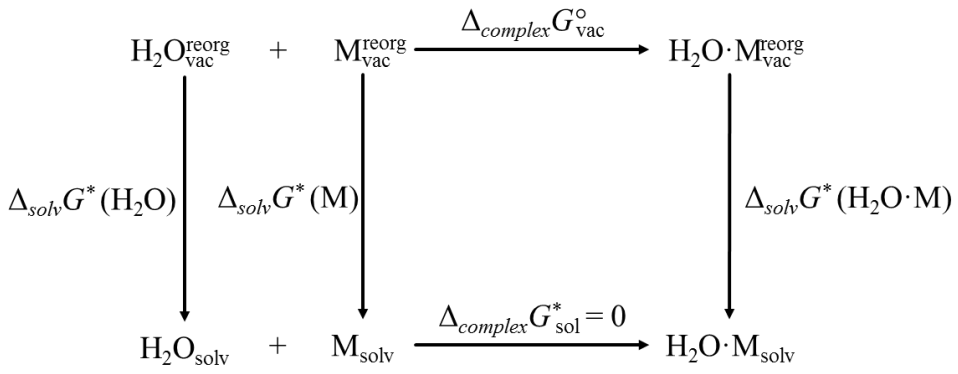


Figure 3: Thermodynamic cycle of the complexation of a molecule M with a water molecule in the gas phase (top) and in aqueous solution (bottom).

3.3 Modelling the effect of solvation *in silico*

Thermodynamically the effect of solvation is modelled by excess chemical potentials. To calculate those, a description of the solvent is needed. This will be called a solvent model.

3.3.1 Fluids

The solvent itself is the most important component when modelling solvation effects. In this work three solvents (Lennard-Jones fluid, Weeks-Chandler-Anderson fluid and water) will be treated. The Lennard-Jones fluid and the Weeks-Chandler-Anderson fluid are simple fluids that serve as model systems. They are used to study the fundamentals of solvation and for a better understanding of the related phenomena. They can be used to identify limitations and advantages of solvent models. Then the gained insight is used to improve the theoretical description of water and especially aqueous solvation free energies.

The **Lennard-Jones fluid** (LJ) is a (very simple) model for noble gases. The LJ fluid is a simple monoatomic solvent without charges or dipole moments. However it includes attractive interactions to model the effect of dispersion.

In the LJ fluid the interaction potential u_{LJ} between two atoms is given by

$$u_{\mathrm{LJ}}(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right). \tag{3.15}$$

This formula describes attractive and repulsive interaction between two particles. The LJ potential is short ranged (in comparison to the Coulomb potential). The distance between the atoms (separation) is r, ε is the well depth of the potential and σ is the distance at which the interaction changes from repulsion to attraction.

The LJ potential asymptotically approaches zero. To speed up the calculation of the interaction potential cutoffs are commonly used. Beyond a certain separation, the interaction potential is ignored. These truncations introduce discontinuities to the calculations.

To avoid them a potential switching can be performed. Here the potential is slowly switched of as the separation reaches the cutoff. In chapter 6.3 the LJ potential in the MD simulations was multiplied which a function interpolating between one at an inner cutoff and zero at the outer cutoff distance[27,28]

$$s_{\mathrm{w}}(r) = \begin{cases} 1 & \text{if } r \le r_i \\ \dfrac{(r_c^2 - r^2)^2 (r_c^2 + 2r^2 - 3r_i^2)}{(r_c^2 - r_i^2 2)^3} & \text{if } r_i < r < r_c \\ 0 & \text{if } r \ge r_c \end{cases}. \tag{3.16}$$

When describing LJ fluids one usually uses reduced units. The separation is expressed in terms of multiples of σ and the energy is given in multiples of ε. Then the state of the LJ fluid is described by the reduced temperature T^* and the reduced number density ρ^*[29]

$$\begin{aligned} T^* &= \frac{k_{\mathrm{B}} T}{\varepsilon}, \\ \rho^* &= \rho \sigma^3. \end{aligned} \tag{3.17}$$

When reduced units are used, every LJ fluid has the same properties independent of the actual values of σ_{LJ} and $\varepsilon_{\mathrm{LJ}}$.

There are equations of state for the LJ fluid (LJEoS) that are purely based on empirical fits. They can be used to calculate thermodynamic properties like the excess chemical potential, the Helmholtz energy, the inner energy and the pressure of the LJ fluid at different liquid states.

One of them is the modified Benedict-Webb-Rubin (MBWR) equation of state[30,31]. This is an equation with 33 adjustable parameters. The MBWR equation is fitted to results from three different simulation techniques of the LJ fluid at different temperatures and densities:

- MD simulations,
- Monte Carlo (MC) simulations,
- particle insertion simulations.

The refined version of the MBWR[31] can predict the Helmholtz free energy, the Gibbs free energy, the pressure, virial coefficients and the inner energy in the range of $0.7 < T^* < 6$. The MBWR is parametrized at all densities that belong to a fluid state within the given temperature range.

The authors[31] performed the simulations with a 4σ cutoff and potential shifting. They applied long range tail corrections to calculate the pressure and inner energy.

This solvent is discussed together with the WCA-rep solvent in chapters 5 and 6.

The **Weeks-Chandler-Anderson** (WCA) fluid can be derived from the LJ fluid by a partitioning of the LJ interaction into an attractive and a repulsive part. The partitioning scheme was developed by Weeks, Chandler and Anderson[32]. This hypothetical fluid is an important system for density functional theory because it allows for perturbation based approaches towards the LJ fluid[33,34,35].

The repulsive potential is given by

$$u_{\mathrm{WCA-rep}}(r) = \begin{cases} u_{\mathrm{LJ}}(r) + \varepsilon & r < 2^{1/6}\sigma \\ 0 & r \geq 2^{1/6}\sigma \end{cases}. \tag{3.18}$$

Consequently the attractive part is

$$u_{\mathrm{WCA-attr}}(r) = \begin{cases} -\varepsilon & r < 2^{1/6}\sigma \\ u_{\mathrm{LJ}}(r) & r \geq 2^{1/6}\sigma \end{cases} = u_{\mathrm{LJ}}(r) - u_{\mathrm{WCA-rep}}(r). \tag{3.19}$$

This solvent is discussed together with the LJ fluid in chapters 5 and 6.

Water is probably the most important solvent on earth. It consists of one oxygen atom and two hydrogen atoms. The hydrogens are quite small in comparison to the oxygen. Water has a large dipole moment and a fairly high relative dielectric constant (ε_r=78). The complex granularity caused by hydrogen bonding complicates the modelling. Since phenomena like hydrogen bonding of hydrophobic hydration are not even fully understood yet, modelling of water is not an easy task.

In this work the (M)SPC/E water model is used. It is a three site water model, which means, that the water molecule is build up from three interaction sites, each centered on one of the atoms. The centers are connected by rigid bonds. The centers interact with the environment using LJ and Coulomb interactions. There are different versions of the SPC/E water model. Here the SPC/E[36] optimized for Ewald summation techniques and the MSPC/E[37,38,39] water model, which was parameterized for integral equation theory are used. The force field parameters and the topology of these modes are summarized in table 1.

For a further discussion of this solvent see chapter 7.

Table 1: Force field parameters and geometry of the SPC/E and MSPC/E water model.

	SPC/E	MSPC/E
q_O	-0.8476	-0.8476
q_H	0.4238	0.4238
σ_O / Å	3.166	3.166
σ_H / Å	0.000	1.000
ε_O / kcal mol-1	0.1553	0.1553
ε_H / kcal mol-1	0.0000	0.0560
H-O-H angle	109.47°	109.47°
O-H distance / Å	1.0	1.0

3.3.2 Solvation models

There are various approaches towards a theoretical description of the solvent. However, they can be assigned to one of two general classes. The first general approach is to surround the solute by a large number of solvent molecules. This is called an explicit solvation model. The members of the other class model the solvent via its (equilibrium) properties. This is called an implicit solvation model. Examples for implicit solvation models are the continuum based approaches which will be shown in chapter 3.5 and the statistical solvent models based on integral equation theory as described in chapter 4.

In an **explicit** solvation model, the solvent is built up from a large number of individual solvent molecules. An ensemble of solvent configurations is generated. This is usually done in molecular dynamics (MD) simulations.

These methods are time consuming because the ensemble of solvent configurations is of infinite size and all the relevant regions have to be sampled. Additionally the number of solvent atoms has to be large enough. Otherwise finite size effects will be introduced. On the other side each molecule added increases the computational burden.

Explicit solvent models allow for the investigation of solvation phenomena with atomic (or even electronic in case of quantum mechanics) resolution. One of the main approximations to the calculations is that the configuration sampling is complete. Therefore a critical part of the error introduced is of a statistical nature and can be minimized by increasing the amount of calculations.

An alternative to explicit solvation models are **implicit** solvation models. They describe the solvent by its macroscopic properties. This accelerates the calculations, since there is no sam-

pling of solvent degrees of freedom and the interactions within the solvent are not needed for these kinds of calculations. But usually there are approximations within such a model that introduce a systematic error.

3.4 The excess chemical potential

The excess chemical potential is a central property for the description of chemical processes in solution. Unfortunately μ^{ex} calculations are not straightforward. The excess chemical potential is an equilibrium property that includes entropic contributions. To calculate these equilibrium properties, averaging over all (or at least "enough") possible solvent configurations is necessary. This introduces either a combinatorial restriction or a sampling component to the calculation or further approximations are needed to cover this part.

Additionally the solvent itself usually is not the focus of research interest. However, in most cases the interaction between a solvent and a solute or the influences of the solvent on a reaction is studied. The number of all solvent atoms is usually much larger than the number of solute atoms. If the solvent is treated in the same way as the solute much computational time is used on the solvent, which is usually the rather uninteresting part of the system.

3.4.1 Databases with experimental excess chemical potentials

As discussed previously, experimental datasets are important for the generation of theoretical models. The quality of a solvation model may be benchmarked against such a dataset, or the model itself may be parameterized with them. Models that supplement theoretical calculation with parameterized quantities from experimental datasets are called semi-empirical.

The Minnesota Solvation Database (MNSol) is a database containing the free energies of solvation. It was compiled by Cramer, Thrular and coworkers[18]. Among others, it consists of over 500 aqueous solvation free energies of small molecules. This has already been used for the parameterization of semi-empirical models such as SM*x*[10,11,12].

3.4.2 Thermodynamic integration

As shown in (3.3) the excess chemical potential is the change of free energy when adding a single particle with a frozen geometry to a system. It can be calculated by integration of all energy contributions along the path of the particle into the solvent. The path integral can be converted into coupling parameter integration. Here the system is extended by hypothetical coupling parameter λ. At the beginning (λ=0) the system is not perturbed. At λ=1 the additional particle is comlpetly added to the system. Along this path the interactions between

system and solute are gradually coupled and consequently the interaction is scaled from zero to the full interaction

$$\mu^{\mathrm{ex}} = \int \mathrm{d}\mathbf{r} \int_0^1 \mathrm{d}\lambda \frac{\mathrm{d}u(\lambda,\mathbf{r})}{\mathrm{d}\lambda} g(\lambda,\mathbf{r}) \,. \tag{3.20}$$

This approach is called thermodynamic integration TI.

3.4.2.1 Thermodynamic integration in molecular dynamics simulations

A thermodynamic integration method is often used in combination with MD simulations. The solvent configurations of an explicit solvent model are sampled along the coupling parameter path. The excess chemical potential can be calculated from the trajectory average of the interaction potential derivative by

$$\mu^{\mathrm{ex}}(\mathbf{r}_{\mathrm{u}}) = \int_0^1 \left\langle \frac{\partial U(\lambda;\mathbf{r}_{\mathrm{v}})}{\partial \lambda} \right\rangle_{\lambda;p,T} \mathrm{d}\lambda \,. \tag{3.21}$$

Usually 10 to 20 simulations with different values of λ are performed. Then the derivative of the complete system interaction potential U with respect to λ is calculated and averaged over the trajectory. The complete system interaction potential U can be approximated by a sum of pair potentials u. The excess chemical potential is given by numerical integration of (3.21).

3.4.2.2 Coupling parameter dependent potentials

For the LJ potential (3.15) the most straightforward way is to multiply the ε parameter with the coupling parameter

$$u_{\mathrm{LJ}}(\lambda, r) = (\lambda\varepsilon)\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right). \tag{3.22}$$

However this leads to very sharp and repulsive potentials at small values of λ. This results in unstable simulation whenever particles come close to each other. To avoid these numerical hurdles one introduces an artificial distance between the scaled particle and the solvent[40]. This can be thought of as if the particle would be introduced from a 4 dimensional hyperspace into the 3 dimensional simulation cell. The fourth dimension is the coupling parameter. This is called soft core scaling.

The Lennard-Jones potential can be extended by a soft core scaling in many different ways. The one chosen here is taken from AMBER 12[41]

$$u_{\mathrm{LJ}}(\lambda, r) = (4\lambda\varepsilon)\left(\left(a(1-\lambda)+\left(\frac{r}{\sigma}\right)^6\right)^{-2} - \left(a(1-\lambda)+\left(\frac{r}{\sigma}\right)^6\right)^{-1}\right). \tag{3.23}$$

The factor $a\lambda$ is the distance between solute and solvent along the fourth dimension. Since the excess chemical potential is a state function it does not depend on the coupling parameter path in equation (3.21). Therefore the a parameter can be chosen freely, but it may influence the stability of the simulation.

For the WCA-rep potential (see chapter 3.3.1) an analogous coupling parameter-dependent representation can be formulated

$$u_{\mathrm{WCA-rep}}(\lambda, r) = \begin{cases} u_{\mathrm{LJ}}(\lambda, r) + \lambda\varepsilon & r < (2-\alpha+\alpha\lambda)^{1/6}\sigma \\ 0 & r \geq (2-\alpha+\alpha\lambda)^{1/6}\sigma \end{cases}. \tag{3.24}$$

However when using soft core scaling, the cutoff distance should be scaled as well. Otherwise an attractive force is introduced into the system and the potential limit for infinite separation will not be zero.

3.4.2.3 Combination of coupling parameter dependent potentials

The interaction potential may consist of two or more additive potential terms. For example an atom could be described by the WCA-rep potential to model Pauli repulsion. Then a WCA-atr potential is added to model dispersions and finally a charge term is added using Coulomb interaction. In the following the handling of those additive potentials in the context of coupling parameter integration is described.

Since the excess chemical potential is a state function, it can be calculated from a multistep process. For each step a partial excess chemical potential is calculated and summed to give the full excess chemical potential. This is necessary when purely attractive potentials are part of the simulation. For every λ and small separations, one has to take care that the repulsive potentials dominate the interaction at close separation. Otherwise the system will contract in an unphysical way. Therefore thermodynamic integrations of systems with Coulomb potentials (e.g. polar solutes in water) are usually treated with two separate simulations. In a first step the excess chemical potential of uncharged system (a pure LJ system) is calculated. The charging free energy is calculated by scaling the Coulomb potential followed by another coupling parameter integration.

In a similar way the free energy of transition from the WCA-rep fluid to the LJ fluid can be calculated with such a coupling parameter-dependent potential

$$u_{\mathrm{LJ}}(\lambda, r) = u_{\text{WCA-rep}}(r) + \lambda u_{\text{WCA-atr}}(r). \tag{3.25}$$

This forms the basis for analytical perturbation theory.

3.5 Dielectric continuum techniques

A popular class of implicit solvent models treats the solvent as a dielectric continuum. The solute is placed into a cavity in the continuum. The interaction potential of the cavity and the solute can be calculated from the Poisson-Boltzmann equation. The potential can be used in the Hamiltonian of quantum mechanics (QM) calculations and MD simulations. In this case the free energy of solvation can be separated into the contributions[42,43,44] of the electrostatic solute-solvent interaction $\Delta_{solv}G_{\mathrm{el}}$, the cavitation formation energy G_{cav}, the dispersive solute-solvent interaction $\Delta_{solv}G_{\mathrm{cav}}$ and the entropic contributions from molecular motion $\Delta_{solv}G_{\mathrm{mm}}$.

$$\Delta_{solv}G = \Delta_{solv}G_{\mathrm{el}} + G_{\mathrm{cav}} + \Delta_{solv}G_{\mathrm{dis}} + \Delta_{solv}G_{\mathrm{mm}}. \tag{3.26}$$

Different versions of this technique exist. They differ for example in the way the Poisson equation is solved, the cavity is formed and how the terms in equation (3.26) are calculated, approximated or neglected.

Commonly used dielectric continuum models are used in MD simulations are the generalized Born model GB[45] and the analytically linearized Poisson Boltzmann ALPB[46] model. Both models are approximations that allow for a fast evaluation of the Poisson-Boltzmann equation. The SM*x* models[47,48,49] use the generalized Born method in combination with quantum mechanics. They allow for a very accurate calculation of solvation free energies and are optimized for a broad spectrum of basis sets and theory levels. All these models (GB and ALPB) use atom centric point charges to model the interaction between the solute and the solvent.

In QM calculations a commonly uses implicit solvation model is the polarizable continuum model (PCM) in one of several variants[43]. These techniques solve the Poisson equation using apparent surface charges. These charges are spread on the surface of the cavity to model solvent induced polarization effects. The PCM variants differ in the way the how the apparent surface charges are calculated.

In chapter 7, the integral equation formalism variant IEF-PCM[50,51,52] of the polarizable continuum model will be used. The IEF-prefix will be omitted. IEF-PCM uses integrals over Green functions instead of discrete sums to calculate the potential energy.

Further widely used variants of PCM models are the conductor like screening model COSMO[53] (also called C-PCM) and its extension COSMO-RS[54] which are mentioned in chapter 7. In COSMO the solvent is modeled as an ideal conductor ($\varepsilon = \infty$). This eases the calculations of the surface charge distribution but the distribution has to be scaled by an empirical correction function to get the charge distribution for a real solvent (that is not an ideal conductor). The quality of the empirical correction gets worse the lower the dielectric constant of the solvent is[43]. COSMO-RS was developed to deal with solvent granularity and directional effects such as H bonds which are needed for a proper treatment of water.

In all models presented so far the cavity in the solvent is calculated from predefined atom radii. In the self-consistent continuum solvation model (SCCS)[55,56] the cavity is defined in an alternentive way. Here an isodensity surface is calculated at a given threshold and everything inside the surface is considered to be inside the cavity. An extension of SCCS is the charge-asymmetric nonlcally determined local-electric solvation model CANDLE[57]. With this model calculation of solvation free energies of ions is more accurate.

4 Integral equation theory

Integral equation based solvation models belong to the class of statistical solvation models. The solute is modeled by its interaction potential with the solvent. The solvent is modelled by its density susceptibility function. This function describes how a fluid's local density changes, when an external potential is applied to the fluid. Integral equation based solvent models account for solvent granularity. As an example RISM is able to model the effect of H-bonds. So it is an implicit solvent model which shows more detail than the dielectric continuum techniques but less than an explicit solvent model would provide. In theory integral equation based solvent models provide the same information about equilibrium properties as an MD simulation.

In integral equation theory the solvent is models by local density derivations using classical density functional theory[58]. Here the thermodynamic properties of a (simple and inhomogeneous) liquid are expressed as a functional of the local density. For example the in the grand canonical ensemble the grand potential Ω is given by

$$\Omega = A - \mu N = A[\rho(\mathbf{r})] - \mu \int \rho(\mathbf{r}) d\mathbf{r}. \tag{4.1}$$

Density functional theory says that the real density minimizes the grand potential. Hence the real density can be calculated by setting the functional derivative of (4.1) equal to zero

$$0 = \left. \frac{\delta \Omega}{\delta \rho(\mathbf{r})} \right|_{\rho = \rho(\mathrm{eq})}. \tag{4.2}$$

From equation (4.2) Ornstein-Zernike OZ equation may be derived using non-trivial algebra[58,59]. For a simple homogenius fluid the OZ is

$$h(r) = c(r) + \rho \int c(r - r') h(r') \mathrm{d}\mathbf{r}'. \tag{4.3}$$

Here r is the spatial vector and $r = |\mathbf{r}|$ is the separation. It relates the total correlation function (TCF or $h(r)$) to the direct correlation function (DCF or $c(r)$). The direct correlation function can be interpreted as the correlation between two particles that is not mediated by other particles. The long range asymptotic of this function is similar to the interaction potential between those particles. The total correlation function is closely related to the pair distribution function (PDF or g(r))

$$h(r) = g(r) - 1 \tag{4.4}$$

where r is the distance. In most cases neither the total nor the direct correlation functions of a system are known. Therefore a second equation connecting h and c is needed. For this purpose the closure relation

$$h(r)+1=\exp[h(r)-c(r)-\beta u(r)+B(r)], \tag{4.5}$$

which is a direct result of equation (4.2)[58,59] is used. Here the correlation functions are linked to the interaction potential $u(r)$ and the bridge function $B(r)$. The correlation functions, bridge function and pair potential of the LJ fluid are shown in figure 4. The closure relation will be discussed in chapter 4.3 and the bridge function will be discussed in chapter 4.4. The inverse temperature β is calculated form the Boltzmann constant k_B and the temperature T by $\beta = 1 / (k_B T)$. Often the indirect correlation function (ICF or $t(r)$) is introduced.

$$t(r)=h(r)-c(r). \tag{4.6}$$

So far the Ornstein-Zernike equation only holds for homogenous and monoatomic solvents. The molecular OZ[60,58] (MOZ) equation is an extension of the OZ equation which allows for a treatment of ridged polyatomic systems. For a solute at a given point in space with a given orientation, integration is performed over all the space coordinates as well as all three angular coordinates describing the relative orientations between solute and solvent. This results in a six dimensional integral[61,62].

4.1 The reference interaction site model

To reduce the computational burden needed for solving the (molecular) Ornstein-Zernike equation the reference interaction site model (RISM)[63,64]was introduced. In contrast to the MOZ, in RISM a factorization of the integration of spatial and angular coordinates is assumed and the angular integration is performed analytically. To do so, the direct correlation function is approximated by a sum of partial site-site direct correlation functions. This isotropic approximation means that the solvent is assumed to rotate freely, ignoring all privileged orientations[58].

This approach leaves a three dimensional integral over the space coordinates which is solved numerically. This numerical solution can be calculated in two different ways called 1D-RISM[65,66] and 3D-RISM[1,67,4].

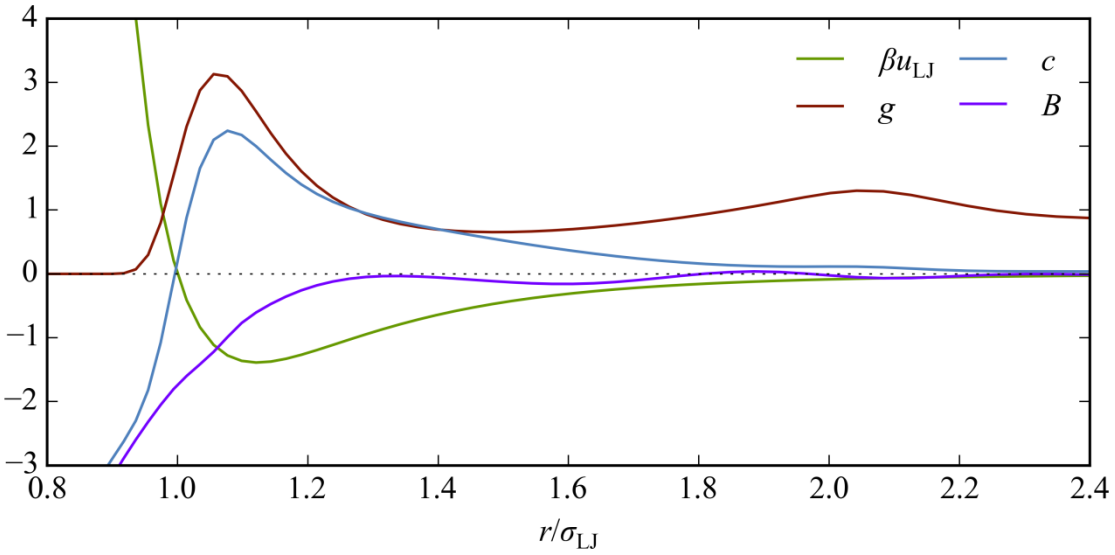


Figure 4: Pair potential u_{LJ}, pair distribution function g, direct correlation function c and bridge function B of the LJ fluid at the reduced density of $\rho^* = 0.85$ and the reduced temperature $T^* = 0.72$. For details of the calculation see chapter 5.

4.1.1 1D-RISM

In the reference interaction site model the solvent is assumed to be built from spherical interaction sites. In the case of a mono atomic fluid the three dimensional integral of the space coordinates reduces to a one dimensional integration for symmetry reasons.

To get the pair distribution function of molecular liquids the solvent is decomposed into spherical components (interaction sites). The geometry of the solvent molecules is modelled by the so called intramolecular correlation function. The intramolecular correlation function is a distribution function that describes the distance distribution of each pair of sites inside the molecule. When the molecule is described as a rigid body of integration sites, the intramolecular correlation function $\omega_{\gamma\gamma'}$ between the solvent interaction sites γ and γ' is given by

$$\omega_{\gamma\gamma'}(r) = \frac{\delta(r - r_{\gamma\gamma'})}{4\pi r_{\gamma\gamma'}^{\,2}}. \tag{4.7}$$

Here δ is the Dirac delta function. For a single solute molecule u with the solute interaction sites α in a molecular solvent v with the solvent interaction sites γ, the integral equation (according to RISM) is

$$h_{\alpha\gamma'}(r) = \sum_{\gamma'}\sum_{\alpha'}\int\int \omega_{\alpha\alpha'}(|\mathbf{r}_1 - \mathbf{r}'|) \times c_{\alpha'\gamma'}(|\mathbf{r}'-\mathbf{r}''|)\chi_{\gamma\gamma'}(|\mathbf{r}''-\mathbf{r}_2|)\mathrm{d}\mathbf{r}'\,\mathrm{d}\mathbf{r}''. \tag{4.8}$$

This is called the 1D-RISM$_{uv}$ equation. It depends on the susceptibility function χ of the pure solvent which is defined as

$$\chi_{\gamma\gamma'}(r) = \rho\omega_{\gamma\gamma'}(r) + \rho h_{\gamma\gamma'}(r). \tag{4.9}$$

The susceptibility can be calculated using the Percus trick. Here the pure solvent is treated as a system of one solvent molecule in the solvent of interest. Then there are no more solute-dependent terms in (4.9) and it can be solved together with a closure relation. This approach is called one dimensional solvent solvent reference interaction site model. (1D-RISM$_{vv}$).

The result of the 1D-RISM$_{vv}$ is the pair distribution function of the pure solvent. This can now be used to calculate the pair distribution function of any solute in this solvent.

For the treatment of large anisotropic molecules, 1D-RISM$_{uv}$ may lead to poor description of local solvent densities near the solvent due to the radial averaging[68]. A spatial (grid) localization of regions with increased (or decreased) density is not possible. Since this information is often required for molecular modelling the 3D-RISM[1,4,67] was developed.

There are two extensions to the RISM formalism. The first one is the extended RISM (XRISM). It was developed by Hirata and Rossky[69]. XRISM is an *ad hoc* generalization of RISM to molecules. The problem of XRISM is that unlike in reality, the dielectric constant ε of every solvent is given by the dielectric constant of a polar the ideal gas[58]

$$\varepsilon = 1 + \frac{4}{3}\pi\beta\rho\mu_{\mathrm{D}}^2. \tag{4.10}$$

Here μ_{D} is the diploe moment of a solvent molecule. To overcome this, the dielectrically consistent RISM[70,71] (DRISM) was developed. Here an additional term is added to the closure relation (similar to a bridge function) that accounts for screening effects.

4.1.2 3D-RISM

3D-RISM[3,4] is an extension of 1D-RISM. Here the spatial distribution of the solvent density is not averaged and information about 3D solvent distribution is gained. In 3D-RISM, for every solvent site γ, the equation (4.8) is partially solved by omission of integration of orientational coordinates

$$h_{\gamma}(\mathbf{r}) = \sum_{\gamma'} \int c_{\gamma'}(|\mathbf{r} - \mathbf{r}'|)\rho^{-1}\chi_{\gamma\gamma'}(|\mathbf{r}'|)\mathrm{d}\mathbf{r}'. \tag{4.11}$$

The solvent susceptibility function can be calculated with 1D-RISM$_{vv}$. To avoid the approximations in the 1D-RISM$_{vv}$ formalism, the solvent susceptibility can be taken from other

sources, for example from pair distribution functions from MD simulations. However, the susceptibility function calculated from external sources keep the approximation that the direct correlation function is built from a sum of site-site direct correlation functions.

4.2 Solute solvent interaction potentials

In equation (4.5) the interaction potential between solute and solvent (or two solvent molecules in the vv case) is needed. This is usually calculated using usual interaction pairwise potentials and force fields. A probe of every solvent site is placed on all positions in the vicinity of the solute and the interaction energy is calculated. In most cases LJ interaction and Coulomb interaction are used. The parameters of the LJ potential are taken from force fields and the partial charges are calculated via quantum mechanics or taken from a force field.

In the EC-RISM[5] approach the partial charges of the solute are calculated in an iterative process. Quantum mechanics and RISM calculations are performed successively. Hence the resulting partial charges account for the solvent induced polarization of the solute.

In EC-RISM the solvation free energy is approximated by

$$\Delta_{slov}G = \Delta_{pol}E + \mu^{\mathrm{ex}} . \tag{3.4}$$

The excess chemical potential is the change in free energy when a rigid molecule with a frozen solution phase electronic structure is inserted at infinite dilution. It is calculated from RISM as explained in chapter 4.5. The quantum mechanical energy difference $\Delta_{pol}E$ is given by

$$\Delta_{pol}E = E^{\mathrm{solv}} - E^{\mathrm{vac}} . \tag{4.12}$$

It can be interpreted as the per-particle polarization energy of a molecule during the solvation process. The energy is evaluated using the solution phase wave function of the molecule. The wave function is calculated using the Hamiltonian

$$\begin{aligned} H_{\mathrm{tot}} &= H_{\mathrm{u}} + H_{\mathrm{uv}} \\ &= (H_{\mathrm{ne}} + H_{\mathrm{ee}} + H_{\mathrm{nn}}) + (H_{\mathrm{nq}} + H_{\mathrm{eq}} + H_{\mathrm{qq}}). \end{aligned} \tag{4.13}$$

Here H_u is the standard molecule Hamiltonian in quantum mechanics and H_{uv} describes the solute-solvent interaction. The subscripts n, e denote nuclei and electrons. The q denotes the embedding point charges modeling the solvent environment. The charge field is calculated from the total correlation functions from 3D-RISM. The density of the charge field ρ_q is given by

$$\rho_q(\mathbf{r}) = \sum_\gamma q_\gamma \rho_\gamma (h_\gamma(\mathbf{r}) + 1) . \tag{4.14}$$

The solvent-site density is given by ρ_γ and the solvent site partial charge is q_γ.

EC-RISM begins with an initial guess of the partial charges of the molecule. This can be done using vacuum QM calculations. Then a RISM calculation is performed and the embedding charge field is calculated using equation (4.14). A QM calculation is performed with the updated Hamiltonian according to equation (4.13). From the QM calculation updated partial charges are calculated and the embedding cluster is recalculated. This is repeated until the solvation free energy of the molecule (3.4) is converged.

4.3 Closure relation with approximated bridge functions

The integral equation shows the response of a solvent after perturbation of the correlation function. Now another equation is needed that models this perturbation introduced by the solute which applies an external potential u to the solvent. This is modeled by the closure relation. This equation can be given in the one-dimensional case as well as in the three-dimensional case.

1D-RISM

$$h(r) + 1 = \exp[h(r) - c(r) - \beta u(r) + B(r)] \tag{4.5}$$

3D-RISM

$$h(\mathbf{r}) + 1 = \exp[h(\mathbf{r}) - c(\mathbf{r}) - \beta u(\mathbf{r}) + B(\mathbf{r})] \tag{4.15}$$

Equation (4.15) can also be used in the 1D-RISM case. Equation (4.5) is identical to (4.15) in 1D-RISM due to symmetry. In the remainder of this chapter all equations are only displayed for the one dimensional case. The three dimensional case is analogous. Here B is a function that is called bridge function due to the formal similarly with the bridge function in classical density functional theory and integral equation theory. This function will be discussed in chapter 4.4. The bridge function B is usually not known, some approximate closures were introduced in the literature. Some of them will be presented in the next chapters.

4.3.1 The hypernetted chain approximation

The approximation that the bridge function can be neglected ($B(r) = 0$) is called the hypernetted chain approximation[13,14] (HNC). The HNC approximation is appropriate for longer ranged interactions since the bridge function vanishes with growing interaction distance. However for real systems like water the HNC leads to unphysical oscillations of the pair distribution function at high separations[72] showing the effect of the missing long range bridge function.

4.3.2 Partial series expansion

The reference interaction site model and hypernetted chain approximation closure system of equations (RISM/HNC) often has convergence issues. To avoid some of them, the closure is partially expressed as a Taylor series[73,74,15].

$$h(r)+1=\begin{cases}\exp(h(r)-c(r)-\beta u(r)) & h(r)-c(r)-\beta u(r)\le 0\\ \sum_k [h(r)-c(r)-\beta u(r)]^k / k! & h(r)-c(r)-\beta u(r)>0\end{cases}. \tag{4.16}$$

Here k is an order parameter of the series expansion. If $k = 1$ the closure is equal to the Kovalenko-Hirata closure (KH) [73,74]. With increasing order, the partial series expansion of order k (PSE-k) approaches the HNC.

4.3.3 Verlet closure

The Verlet closure[75] is a semi-empirical closure relation. It was developed to reproduce the pressure of hard sphere systems. Several variants of this closure exist. Here two of them are presented which will be relevant for this work.

The first variant of the Verlet closure is the original form of the closure. It can be used for solutes without long range interactions

$$B_V(r)=-\frac{1}{2}\left(\frac{t^2(r)}{1+0.8t(r)}\right). \tag{4.17}$$

When long ranged interactions are present, (e.g. because the solvent is charged) a renormalization of the indirect correlation function t is needed[76,77]. A possible modified Verlet closure including a renormalization (see chapters 4.4.1 and 4.5) is [78]

$$B_{VM}(r)=-\frac{\xi}{2}\left(\frac{t^*(r)^2}{1+0.8t^*(r)}\right). \tag{4.18}$$

Here $t^*=t-\beta u_{\text{WCA-atr}}$ is the renormalized indirect correlation function and ξ is an adjustable parameter.

4.3.4 Reference hypernetted chain approximation

When the bridge function of a similar fluid is known, instead of the exact bridge function of the fluid of interest, the reference HNC closure can be used. Here the solvent is modelled as a solvent with the artificial potential $u^{ref}=u+B^{ref}$ where u is the potential energy of the solvent and B^{ref} is the bridge function of the reference fluid. With this reference potential the HNC closure is solved[79]. As a reference fluid the hard sphere fluid can be used since its bridge function is known from Monte Carlo simulations[80].

An example for such an approach is the RISM/HNC-BF0 theory[81], where a two-step process is used. In the first strep, the charges solvent are removed. Then the RISM/Verlet approximation is used to calculate B_V of the resulting short ranged system. In the second step, the bridge function of the charged system is approximated by the bridge function of the uncharged system.

4.4 Bridge function

The bridge function is one of the central missing links in integral equation theory. It is a non-local usually unknown functional. When using diagrammatical expressions within integral equation theory, the bridge function is given by an infinite sum of high dimensional integrals which usually does not converge after the first few elements[82,83]. In practice this approach towards the bridge function cannot be utilized in reasonable computational time. Therefore usually closures with approximate bridge functions are used. In this chapter the properties of these bridge functions and other approaches towards its calculation are discussed.

4.4.1 Properties of the bridge function

For a simple liquid the bridge function has a proper representation when described as a function of the separation. However this representation is usually not useful in RISM because of two reasons. Firstly the bridge is rarely known (see chapter 5.1) and hard to approximate in this representation. An example of this representation is the repulsive bridge correction[84]. Secondly this representation cannot be used for the calculation of thermodynamics as shown in chapters 4.5 and 6. To avoid the first issue, the bridge function can be approximated as a function of correlation functions (h, c, t). In this case the explicit dependence on the separation of the correlation functions and the bridge function is used to generate a parametric representation. At every separation the value of the bridge function is assigned to the value of a correlation function. Such an approximation allows for a relatively simple handling of the

bridge function in the RISM[76]. The Verlet closure, which was discussed in chapter 4.3.3, is a good example for a parametric closure.

In the RISM formalism, the bridge function is not correct in the sense of classical functional theory. The RISM site-site bridge function does not minimize the grand potential. However, the bridge function is formulated *ad hoc* in an analogous way to classical functional theory.

Duh and Haymet[76] chose the indirect correlation function as a basis for the parametric representation. The graph of this representation is called a Duh Haymet plot. When the bridge function of the LJ fluid is plotted as a function of the indirect correlation function, the Duh Haymet plot has loops at high separation and hooks at close separation[85]. Hence this representation is not a function but a functional representation of the bridge function. At the example of a 2-2 electrolyte Duh and Haymet[76] proved that the loops in the representation can be avoided when a renormalization is done. Here a specifically chosen long range potential is subtracted from the direct correlation function. When this is done in the correct way[86] this does not affect the result of the RISM but can speed up the calculation.

S. Kast[16] found an upper limit of the bridge function when it is represented as a function of the renormalized direct correlation function $c^* = c + \beta u$. This is a consequence of the Lambert W function and hence a mathematical constraint to the bridge function. The Lambert W condition says that the bridge function is greater than the renormalized direct correlation function. If the total correlation function is zero, the bridge function is exactly the renormalized direct correlation function.

The bridge function is important for the calculation of excess chemical potentials with RISM. This is discussed in the chapters 4.5 and 6.

4.4.2 Non-diagrammatic definition of the bridge function

One can interpret the bridge function as the part of the potential of mean forces W or PMF that is not captured by the direct interaction u or any of the correlation functions (h, c, t).

When the closure equation (4.5) is divided by the Boltzmann weighted interaction (βu), one gets the cavity distribution function[58] γ. Similar to the pair distribution function g, the cavity distribution function describes the distribution between two non-interacting cavities in the solvent. The cavities are not allowed to move within the solvent and the solvent atoms cannot be placed inside the cavity. However the cavities may overlap

$$\gamma(r) = \exp[h(r) - c(r) + B(r)] = \exp[-\beta W^{v}]. \qquad (4.19)$$

here W^{v} is the solvent component of the so called potential of mean forces PMF . Hence the mean force F^{v} on a cavity (or molecule[87]) in a distance r to another cavity (or molecule) is given by

$$F^{v} = -\frac{\mathrm{d}W^{v}}{\mathrm{d}r}. \tag{4.20}$$

F^{v} only measures the attraction or repulsion on the cavity that is introduced by the solvent itself. For a simple fluid, the potential of mean forces W between two molecules at a distance r can be calculated from W^{V} and their interaction pair potential[58] from

$$W(r) = W^{v}(r) + u(r). \tag{4.21}$$

The potential of mean forces is given by

$$\beta W = \beta u(r) + \beta W^{v}(r) = c(r) - h(r) - B(r) + \beta u(r). \tag{4.22}$$

This allows for the calculation of the bridge function

$$B(r) = -\beta W + c(r) - h(r) + \beta u(r) = -\beta W - t^{*}(r). \tag{4.23}$$

4.4.3 Calculation of bridge functions

The approach based on the potential of mean forces from chapter 4.4.2 can be used to calculate bridge functions. Here an inversion of the closure is performed

$$B(r) = \ln[g(r)] + c(r) - h(r) + \beta u(r). \tag{4.24}$$

Commonly the pair distribution function is calculated from MD simulations, but with the problem that the area inside the core of the solute atoms is hardly computable in a straightforward manner. Here the distribution function is (almost) zero and the bridge function is superseded by the repulsive potential. Therefore using equation (4.24) the bridge function can only be calculated outside the core of the solute. As an example for the LJ fluid, the pair distribution function is non-zero roughly when $r > \sigma_{LJ}$. Therefore when using MD simulations, the bridge function can only be calculated easily in this area.

Equation (4.24) also contains the direct correlation function. This function can be calculated from the pair distribution function by an inversion of either the Ornstein-Zernike equation (4.3) or the RISM equations (4.8) or (4.11)[72,80,88,85,89].

To calculate the bridge function inside the core, the cavity correlation function γ or the solvent mediated potential of mean forces W^{v} can be used.

The bridge function of some fluids had been calculated inside and outside the core in the past. One of them is the bridge function of the LJ fluid. It was calculated inside the core by Llano-

Restrepo and Chapman[89]. They used a Monte-Carlo simulation to calculate the pair distribution function and the cavity correlation function. The bridge function was calculated similarly to equation (4.19)[1].

A similar approach can be used with constrained dynamics simulations. Here a dimer of two solvent atoms with a fix distance is treated within a MD simulation. The projection of the mean solvent mediated force along the binding axis can be calculated from [87]

$$F^{\nu}(r_{12}) = \frac{1}{2}\left\langle \frac{\mathbf{r}_{12}}{r_{12}}\left(\mathbf{F}^{\nu}(\mathbf{r}_1) - \mathbf{F}^{\nu}(\mathbf{r}_2)\right)\right\rangle_{r_{12}} . \tag{4.25}$$

Here r_{12} is the distance between the two atoms of the solvent dimer. The position of the dimer atoms is given by $\mathbf{r}$ and the bracket means the trajectory average. Then the solvent mediated potential of mean forces is given via integration of the forces

$$W^{\nu}(r) = -\int_{\infty}^{r} \mathrm{d}r' F^{\nu}(r') . \tag{4.26}$$

Note that the upper integration limit is problematic since we can only treat a limited range of constraint distances by MD simulations. By splitting W^{ν} into

$$W^{\nu}(r) = -\int_{r_x}^{r} \mathrm{d}r' F^{\nu}(r') - \int_{\infty}^{r_x} \mathrm{d}r' F^{\nu}(r') . \tag{4.27}$$

with variable r_x we can circumvent this issue by explicit PMF simulation of the first term on the right hand site and taking the second term directly from the simulated PDF of the pure liquid by

$$-\int_{\infty}^{r_x} \mathrm{d}r' F^{\nu}(r') = -\beta^{-1} \ln g(r_x) - u(r_x) . \tag{4.28}$$

The bridge function is then obtained by supplementing

$$-\beta W^{\nu}(r) = g(r) - 1 - c(r) + B(r) = t(r) + B(r) . \tag{4.29}$$

with an direct correlation function from inverting the Ornstein-Zernike equation.

4.5 Thermodynamics

As shown in chapter 3.4.2 the excess chemical potential can be calculated from a thermodynamic integration. All properties in equation (3.20) (u and g) are either input or the result of a RISM calculation. Similar to MD simulations, RISM calculations can be performed at different values of the coupling parameter λ, followed by a numerical integration. This was for ex-

[1] The following part is taken from Tomazic, D.; Hoffgaard, F.; Kast, S. *Chemical Physics Letters* 2014, *591*, 237-242.

ample done by Chiles and Rossky[90] to calculate the potential energy surface of an S_N2 reaction in water.

Singer and Chandler derived a closed (without an infinite sum) and direct (without coupling parameter integration) expression for the excess chemical potential for RISM calculations with the HNC closure[91]. An analytical integration of (3.20) for the HNC closure leads to

$$\mu^{\mathrm{HNC}} = \frac{4\pi\rho}{\beta}\int r^{r}\left(\frac{1}{2}h^2(r) - c(r) - \frac{1}{2}h(r)c(r)\right)\mathrm{d}r \tag{4.30}$$

and

$$A^{\mathrm{HNC}} = \frac{2\pi\rho}{\beta}\int r^2\left(\frac{1}{2}h^2(r) - c(r)\right)\mathrm{d}r + \frac{1}{4\pi^2\beta\rho}\int k^2\left(\rho\hat{c}(k) + \log[1-\rho\hat{c}(k)]\right)\mathrm{d}k\,. \tag{4.31}$$

In the following chapters equations (4.30) and (4.31) will be referred as the HNC-functionals of the excess chemical potential μ^{HNC} and the Helmholtz free energy A^{HNC}. These functionals may be evaluated with correlation functions calculated from non-zero bridge functions. As an example the HNC-functional excess chemical potential of a RISM calculation with the modified Verlet closure is

$$\mu^{\mathrm{HNC}}[h_{\mathrm{VM}}, c_{\mathrm{VM}}] = \frac{4\pi\rho}{\beta}\int r^2\left(\frac{1}{2}{h_{\mathrm{VM}}}^2(r) - c_{\mathrm{VM}}(r) - \frac{1}{2}h_{\mathrm{VM}}(r)c_{\mathrm{VM}}(r)\right)\mathrm{d}r\,. \tag{4.32}$$

In the following equations μ^{HNC} will be used to describe the HNC part in the general closed form expression of the excess chemical potential. In these equations the functionals are evaluated with the correlation functions calculated with bridge functions.

If a non-zero bridge function is used, the bridge function contributes to the excess chemical potential. This means that the functional for excess chemical potential has two parts, the bridge dependent functional $\Delta\mu^B$ and the HNC functional μ^{HNC}.

$$\mu^B[h_{\mathrm{B}}, c_{\mathrm{B}}] = \mu^{\mathrm{HNC}}[h_{\mathrm{B}}, c_{\mathrm{B}}] + \Delta\mu^B\,. \tag{4.33}$$

The bridge dependent functional is given by[92,39]

$$\begin{aligned}\Delta\mu^B[h_{\mathrm{B}}, c_{\mathrm{B}}] &= \frac{4\pi\rho}{\beta}\int r^2\int\left(B + h_{\mathrm{B}}\frac{\partial B}{\partial\lambda}\right)\mathrm{d}\lambda\mathrm{d}r \\ &= \frac{4\pi\rho}{\beta}\int r^2\int(h_{\mathrm{B}}+1)\frac{\partial B}{\partial\lambda}\mathrm{d}\lambda\mathrm{d}r\end{aligned}\,. \tag{4.34}$$

The bridge dependent functional can have a different expression dependent of the formulation of the bridge function. In this case the functional can be calculated from the bridge function and other correlation functions.

Unfortunately certain closure relations will introduce a path dependence into equation (3.20)[92]. This leads to excess chemical potentials that are no longer state functions. This is called thermodynamic inconsistence. It is possible to define criterions for the bridge function which will lead to thermodynamically consistent excess chemical potentials[78,93,94]. According to S. Kast[78] a bridge function has to be defined in dependence of either t-βu, h or c+βu or certain combinations thereof to get rid of the path dependence.

The identification of path independent bridge functions is the result of a general formalism[78]. The sufficient condition for path independency is the existence of an exact differential of the excess chemical potential hence the variation vanishes

$$\frac{\delta\mu}{\delta h(\mathbf{r},\lambda)} = \frac{\delta\mu}{\delta c(\mathbf{r},\lambda)} = \frac{\delta\mu}{\delta t(\mathbf{r},\lambda)} = \frac{\delta\mu}{\delta u(\mathbf{r},\lambda)} = 0\,. \tag{4.35}$$

In this case the excess chemical potential is given by

$$\mu^{\mathrm{ex}} = 4\pi\int r^2 \int \rho g(r,\lambda)\frac{\partial u(r,\lambda)}{\partial\lambda}\mathrm{d}\lambda\mathrm{d}r + C_L\,. \tag{4.36}$$

C_L are Lagrange constraints imposed by the closure (equation (4.5)), the definition of the indirect correlation function (equation (4.6)) and the integral equation (equation (4.3)).

$$\begin{aligned} C_L = {} & 4\pi\int r^2 \int p(r,\lambda)(\exp[t-\beta u+B]-g) + \nu(r,\lambda)(h-c-t)\mathrm{d}\lambda\mathrm{d}r \\ & + \frac{q}{2\pi^2}\int k^2 \int \hat{c}(k,\lambda)\frac{\partial\hat{c}(k,\lambda)}{\partial\lambda}\hat{\chi}(k) - \rho\hat{h}(k,\lambda)\frac{\partial\hat{c}(k,\lambda)}{\partial\lambda}\mathrm{d}\lambda\mathrm{d}k. \end{aligned} \tag{4.37}$$

In this equation p, ν and q are Lagrange parameters and the r and λ dependency of t, h, c, u, B, and g is left out for the sake of readability. The closure dependent Lagrange parameters are accessible by calculation of the partial derivatives of equation (4.36) and setting them to zero according to equation (4.35).

Then the solvation free energy is calculated using the partial derivative of μ with respect to h.

$$\frac{\delta\mu^{\mathrm{ex}}}{\delta h} = \rho\frac{\partial u}{\partial\lambda} - p - q\rho\frac{\partial c}{\partial\lambda} + \nu \tag{4.38}$$

When equation (4.38) is multiplied with (h+1) the integral of the first term of the right hand side is equal the excess chemical potential when compared with equation (3.20)

$$\begin{aligned} \mu^{\mathrm{ex}} &= \int\int_0^1 \rho\frac{\partial u}{\partial\lambda}(h+1)\mathrm{d}\lambda\mathrm{d}r \\ &= \int\int_0^1 p(h+1) + q\rho\frac{\partial c}{\partial\lambda}(h+1) + \nu(h+1)\mathrm{d}\lambda\mathrm{d}r. \end{aligned} \tag{4.39}$$

Kast and Kloss[15] derived a closed form expression for bridge functions given as a function of the renormalized indirect correlation function t^*

$$\mu^{B(t^*)}[h_B, c_B] = \mu^{HNC}[h_B, c_B] + \frac{4\pi\rho}{\beta}\int r^2 \left(h_B(r) - \int_0^{t^*} dt^*(h_B(t_B^*; r) + 1) \right) dr \ . \qquad (4.40)$$

Using this equation they derived the excess chemical potential for the PSE closures

$$\begin{aligned} \mu^{PSE\text{-}k} &= \mu^{HNC}[h_{PSE\text{-}k} + c_{PSE\text{-}k}] + \frac{4\pi\rho}{\beta}\int r^2 \Theta(h_{PSE\text{-}k}(r))(t^*_{PSE\text{-}k}(r))^{k+1}/(k+1)!dr \\ &= \mu^{HNC} + \Delta\mu^{PSE\text{-}k}. \end{aligned} \qquad (4.41)$$

Here Θ is the Heaviside step function.

An alternative way to express the bridge dependent functional is the $g(B)$ integration which is the result of integration by substitution of equation (4.34)

$$\mu[g(B)] = \mu^{HNC} + \frac{4\pi\rho}{\beta}\int r^2 \int_0^B g(B) dB\, dr \ . \qquad (4.42)$$

The partial molar volume V_M of the solute can be calculated from combination[95,96] of IE theory with Kirkwood-Buff theory[97]. Kirkwood-Buff theory expresses thermodynamic properties in terms of integrals of correlation functions. For the two-component system of a solute particle u in a simple solvent v at infinite dilution the partial molar volume of the solute is given by

$$V_M = \frac{4\pi\kappa}{\beta}\int r^2 (1 - \rho c_{uv}(r)) dr \ . \qquad (4.43)$$

Here κ is the compressibility of the solvent.

5 Bridge function of simple fluids

In chapter 4.4 the general properties and possible ways towards the calculations of bridge functions were explained. Anyways the bridge function is so far only known for few simple fluids. In this chapter the calculation of the bridge function of the WCA-rep fluid is presented and the results are discussed. Additionally this bridge function is compared to the bridge function of the LJ fluid published by Llano-Restrepo and Chapman[89]:

The interest in this context is fourfold[2]:

- tabulating accurate data including the inner-core region in comparison with the full LJ case for possible use by the IE community,
- establishing direct molecular dynamics (MD) simulations of the potential of mean force (PMF) between constrained atom pairs [87] as a means for treating the difficult region,
- comparing the established direct (reciprocal space) inversion of the OZ equation as e.g. used in [98,72] with an iterative real space variant [99,100,101],
- analyzing the parametric dependence of the thus obtained bridge function on a renormalized ICF $t^*=t-\beta u$, which has been proven to be a possible candidate for a thermodynamically consistent bridge function parametrization[78,15]. In particular, such a functional form satisfies the state function condition of the free energy to be independent of the coupling parameter integration path. This will (in principle) allow for an accurate calculation of solvation thermodynamics.

5.1 Fluids with known bridge function

The bridge function of the LJ fluid and the hard sphere fluid are known within the core and outside[89,80]. The bridge function of LJ mixtures[85] is known as well.

For a fluid consisting of dipolar sticks, the angular dependency of the relative orientation of two dipoles on the bridge function is known[102].

For water, methanol and ethanol the site-site bridge functions are known outside of the core only[103,104,72].

[2] The following part and parts of chapter 5.2 are taken from Tomazic, D.; Hoffgaard, F.; Kast, S. *Chemical Physics Letters* 2014, *591*, 237-242.

5.2 The bridge function of the WCA-rep fluid

The bridge function of the WCA-rep fluid was calculated using the constrained dimer method described in chapter 4.4.3. In equation (4.29) the direct correlation function is needed. It is calculated by an inversion of the OZ equation. In case of the repulsive WCA fluid two approaches are possible. These are the direct inversion and the iterative inversion.

Direct inversion

In contrast to molecular site IEs and in the absence of long range interactions the direct inversion of the OZ equation in reciprocal space is unproblematic[98]. Starting with the Fourier transform (FT) of the TCF (assumed to be known over a sufficiently long range where it has practically decayed to zero),

$$\hat{h}(k) = \mathrm{FT}\left[g^{\mathrm{MD}}(r) - 1\right] \tag{5.1}$$

the ICF is given by

$$t(r) = g^{\mathrm{MD}}(r) - 1 - c(r) = g^{\mathrm{MD}}(r) - 1 - \mathrm{FT}^{-1}\left[\frac{\hat{h}(k)}{1 + \rho\hat{h}(k)}\right]. \tag{5.2}$$

The resulting ICF enters (4.29) together with simulated PMF data in order to obtain *B* including the inner core region.

Iterative inversion

The alternative is an iterative, real space inversion[99, 100, 101] based on

$$B(r) = f(r)\left(\ln(g^{\mathrm{MD}}(r)) + \beta u(r) - t(r)\right) \tag{5.3}$$

which, upon insertion in (4.5), basically leads to a constraint on the solution of the OZ / closure system in that *g* is set to the MD values over a specified range. Here, *f*(*r*) defines a switching function (cubic polynomial) varying between 0 and 1 in a transition region beyond which the HNC approximation is assumed to be valid. In practice this means that the OZ equation is solved together with (5.3) as effective closure relation, implying a grand canonical long range correction to the PDF. If *B* would be extracted in this way, the core region would be undefined due to the reasons discussed above. Here, we use this approach solely in order to obtain the ICF, which is then inserted similar to the direct approach together with simulated PMF data for extracting *B* via (4.29).

5.2.1 Computational details

MD simulations were performed with LAMMPS [105] in the canonical ensemble using the Nosé-Hoover thermostat [106,107] and a time step of 5 fs for a range of states corresponding to [89]. LJ parameters were σ = 2.79 Å and ε = 0.0709 kcal mol^{-1} resembling a liquid neon model[108]

(note that all results are specified in reduced units, i.e. for density $\rho^* = \rho\sigma^3$ and temperature $T^* = k_B T/\varepsilon$). Initially, the atoms were placed randomly inside the simulation volume and then preequilibrated in the microcanonical ensemble starting with zero velocities by limiting the maximum displacement per time step to 0.1 Å for 2.5 ps. Similar to the simulation setup in[89] the LJ potential was smoothly switched to zero using a cubic polynomial between 8 and 10 Å for production purposes unless specified otherwise. For reference PDF data (used for direct and iterative inversion as well as for PMF long range offset determination) and for explicit PMF simulations two different setups were used. The former was calculated from production runs based on 15000 atoms; for the latter, the mean forces were recorded for two constrained atoms out of a total of 5002 (to check convergence, for a single state in addition 15000 solvent atoms were simulated) employing the SHAKE algorithm[109].

PDF data was computed after 5 ps microcanonical and 250 ps canonical ensemble equilibration periods from simulations over 5 ns using 10000 snapshots up to a distance of 30 Å with a bin size of 0.06 Å. Since the tail of the PDF oscillates around one for distances between 20 Å and 30 Å no long range correction of the PDF was applied. To estimate the error we calculated for a single one state ($T^* = 1.0$ $\rho^* = 0.8$) the PDF standard error for every 100 snapshots, finding 0.004 in the area of the first peak and around 0.0005 between 20 Å and 30 Å. The resulting data were therefore used without further smoothing. For PMF calculations, 195 distances between 0.01 Å and 19.9 Å were evaluated in steps of 0.03 Å for (0.01-0.1) Å, 0.05 Å for (0.1-6) Å, 0.1 Å for (6-10) Å, 0.3 Å for (10-19.9) Å. 25000 snapshots of 12.5 ns production simulation were evaluated after a 117.5 ps equilibration phase in the canonical ensemble. The mean statistical error of the PMF for all states was found to be below 0.0005 kcal mol^{-1}. For the long range PMF offset, the forces were integrated for 15 different values of r_x between 4.4 and 10 Å followed by averaging over the results. To check the accuracy of the calculations we also derived the PDF from the PMF and determined the maximum deviation from the directly simulated PDF. The largest value for all state points was 0.06; the maximum mean deviation was about 0.004.

The IE calculations were performed on a logarithmic grid of 512 points ranging from $5.98 \cdot 10^{-3}$ Å to a maximum distance of 164.02 Å for the same densities and temperatures as in the MD simulations, employing the Talman method for the fast Fourier transform[110] including zero padding over twice as many points. The IE solutions were converged by the modified direct inversion of iterative subspace (MDIIS) method[111] to a threshold of $\max(\Delta c) < 10^{-5}$ between two successive iteration steps. For the iterative inversion of the OZ / closure equation

the switch function in (5.3) turned off the reference PDF between 17 and 21 Å. The direct inversion of the simulated PDF was performed on a linear grid from 0 Å to 19 Å with a spacing of 0.1 Å. The simulated PDF was interpolated to the target grids using cubic spline interpolation whereby it was set to 1 for $r > 30$ Å. The iterative inversion requires a potential energy which was chosen as identical to the one applied in the MD simulations. For testing the extracted bridge functions they were truncated at 7 Å before reinserting into the OZ/closure equations. Here, the full LJ potential was applied in order to allow for a comparison with reference LJ data.

5.2.2 Analysis of the methodology

Figure 5 illustrates the solvent-mediated component of the PMF for both, the full LJ and the repulsive part of the WCA potential (WCA-rep). The statistical noise is apparently negligible as expected from the size of the simulation systems. Overall, the shape of the PMF curves in the LJ and the WCA-rep case are similar while larger discrepancies are found for the core region; the long range part of W^{ν} is apparently only marginally influenced by the presence of a long range attractive interaction tail. For each state the close contact value of W^{ν} is slightly more attractive for WCA-rep than for LJ while the intermediate range at slightly larger distances than the potential minimum (at around 1.12 σ) is more repulsive for LJ than for WCA-rep.

The two variants for bridge extraction, direct reciprocal space and iterative real space inversion were tested for the LJ fluid in direct comparison with reference data by Llano-Restrepo and Chapman[89] as illustrated in Figure 6. Overall there is a close agreement between the various approaches with the largest discrepancies in the depletion zone between the first and the second RDF maximum at around $r/\sigma = 1.5$. For certain thermodynamic states, larger discrepancies between our and reference data are found in the second depletion region. These differences are of a similar order of magnitude as those found by varying the simulation particle number and cutoff distances, as shown in Figure 7. In general, the cutoff effect is more important than the finite size effect for our simulations; 5000 solvent particles are apparently sufficient for production purposes. The reference data[89] was generated based on a much smaller simulation system which might explain why our bridge data are systematically slightly higher in the core region. The increase of statistical error with larger pair distance is apparently responsible for the increasing discrepancies at longer range.

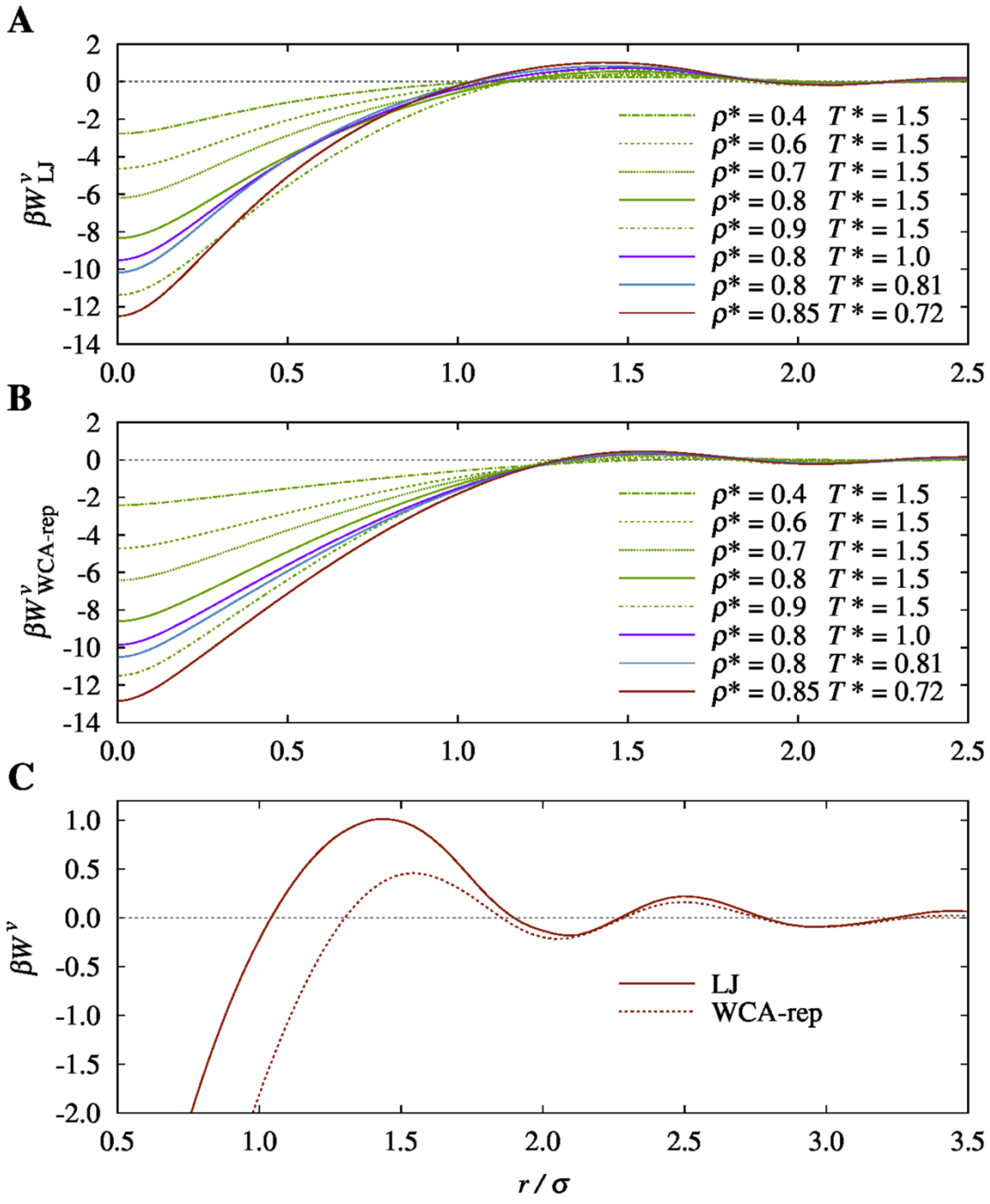


Figure 5 Solvation contribution to the PMF as a function of pair distance for various states represented by different line styles as specified in the panels; **A** LJ fluid, **B** repulsive WCA fluid, **C** enlarged comparison of LJ and repulsive WCA results for $\rho^* = 0.85$ and $T^* = 0.72$.

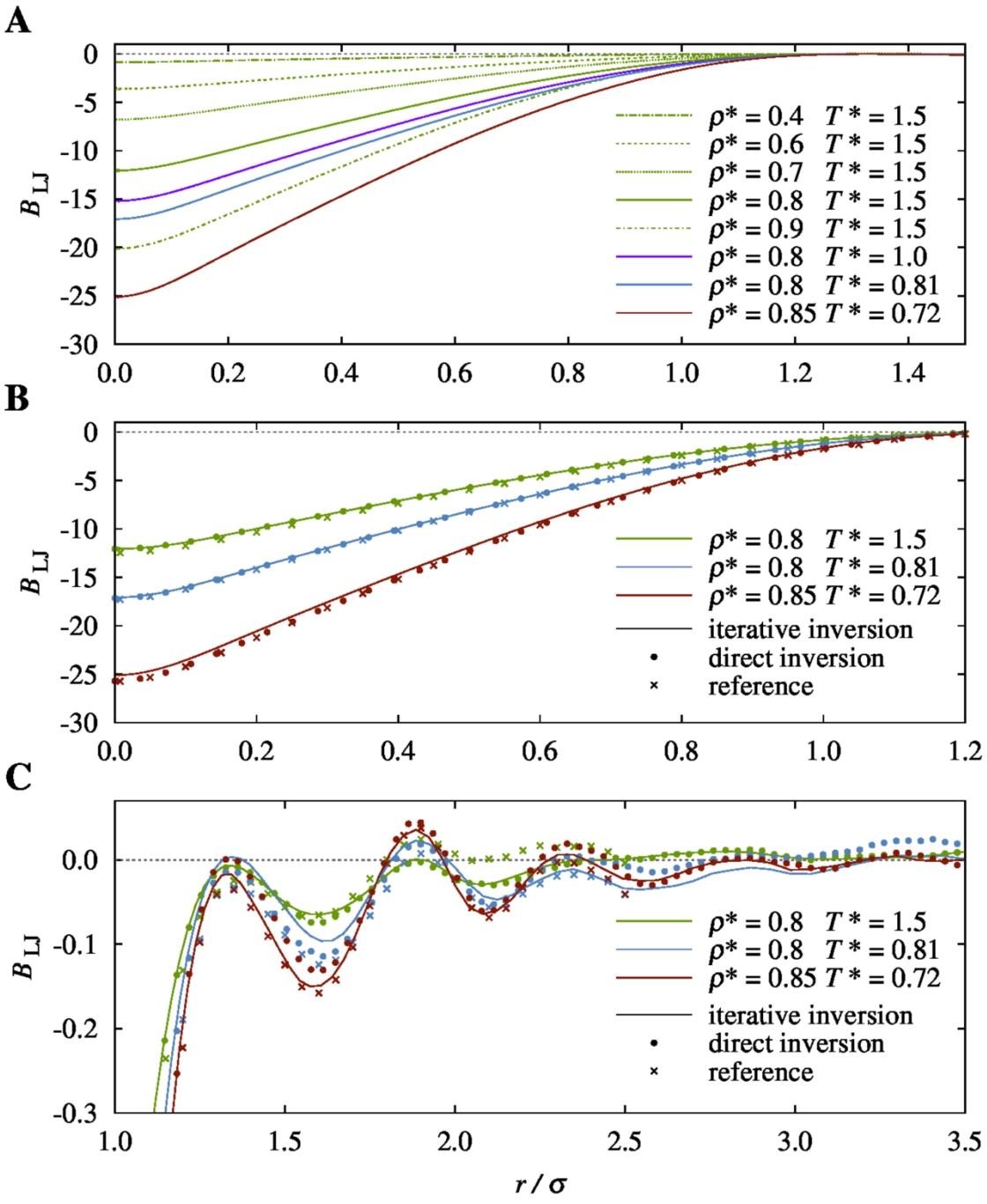


Figure 6 Bridge data for the LJ fluid as a function of pair distance for various states represented by different line styles as specified in the panels; **A** from iterative inversion, **B**,**C** comparison of iterative (real space) and direct (reciprocal space) inversion, and with reference data by Llano-Restrepo and Chapman[89].

The small local differences do not significantly influence the internal energy as derived from IE-computed PDFs obtained by inserting bridge data directly into the closure. Results are summarized in Table 1 where we also compare with data from an analytical equation of state of the LJ fluid[31]. In summary, the iterative and the direct inversion methods are practically

equivalent and reliable. In practice, the iterative variant is simpler to apply to more complex systems described by site-site IEs since it avoids problems of ill-conditioned matrices [72].

Table 2 Reduced excess internal energy $U_{red} = U^{ex}/(N\varepsilon)$ of the LJ fluid for various states obtained from solving the OZ/closure equation with bridge data from different sources, iterative ("iter") and direct ("direct") inversion, reference data by Llano-Restrepo and Chapman[89] ("ref") in comparison with results from the analytical LJ equation of state ("EoS")[31].

ρ^*	T^*	U_{red}^{EoS}	U_{red}^{iter}	U_{red}^{direct}	U_{red}^{ref}
0.4	1.5	-2.70892	-2.70534	-2.70356	-2.69072
0.6	1.5	-3.96024	-3.96508	-3.96128	-3.97364
0.7	1.5	-4.58747	-4.57796	-4.57650	-4.57259
0.8	1.5	-5.13638	-5.11805	-5.11622	-5.13568
0.9	1.5	-5.52627	-5.50659	-5.50820	-5.48975
0.8	1.0	-5.52345	-5.53666	-5.53153	-5.52463
0.8	0.81	-5.71954	-5.70218	-5.70468	-5.70838
0.85	0.72	-6.12489	-6.11456	-6.11788	-6.11703

5.2.3 Analysis of bridge functions

Figure 8 shows bridge data from iterative inversion for the WCA-rep in comparison with the LJ case. Overall, the shapes are quite similar, more so for smaller densities and higher temperatures as expected. The limiting value of *B* for vanishing distance is systematically smaller for LJ compared to WCA-rep. Most significant differences are again found in the first depletion region.

The key property of bridge functions that we want to examine in this work is related to its parametrization as a function of correlation functions. Since the work by Duh and Haymet[33] it is well known that no universal bridge parametrization by the ICF alone exists independent of the thermodynamic state. To this end they established a parametric representation of *B* as a function of *t* for corresponding distances ("Duh-Haymet plot"). On the other hand, as shown earlier[78,15] a bridge function that depends on a renormalized ICF $t^*=t-\beta u$ belongs to a class of bridge models that satisfies the criterion of path independence of the coupling parameter integration leading to the free energy. We have demonstrated this property numerically by explicit integration of a modified Verlet (VM) fit to LJ reference bridge data[78] where the ICF was renormalized by the long range WCA component only. Such an approach is equivalent to scaling the interaction between the WCA-rep and LJ cases studied in this work. It turned out

that good results are obtained in this way which raises the question as to why this is the case. Now numerical evidence is provided since as expected a – at least state-specific – universal bridge parametrization between the WCA-rep and the LJ potentials for the latter approach is given.

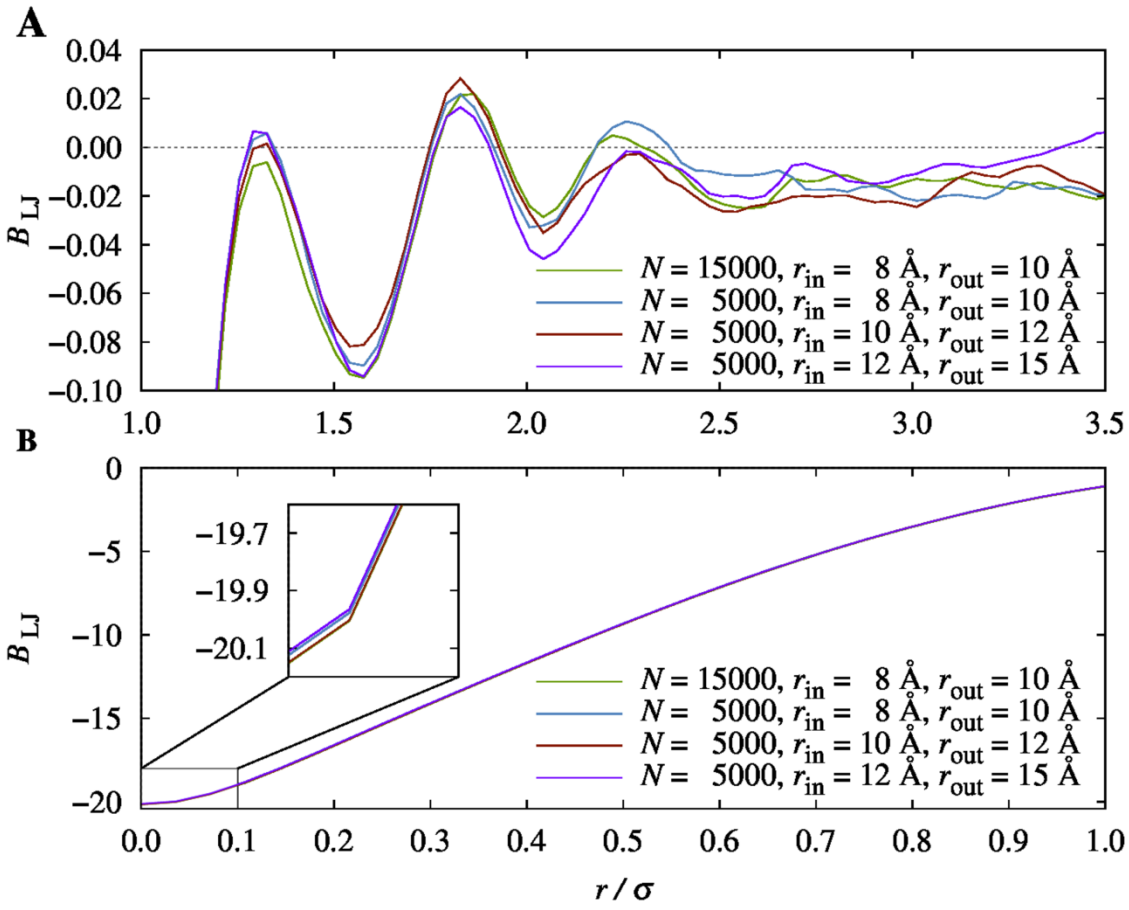


Figure 7 Bridge data for the LJ fluid from direct inversion under various simulation conditions (different particle numbers N and inner/outer ranges r_{in}/r_{out} of the switching function that turns off interactions) as a function of pair distance for $\rho^* = 0.9$ and $T^* = 1.5$; **A**,**B** represent different plot ranges. An 8/10/12/15 Å distance in the real simulated system corresponds to reduced distances r/σ of about 2.87/3.58/4.30/5.38. In the inset, the orange line corresponding to the larger system is indistinguishable from the lowest curve.

Figure 9 depicts the modified (i.e. renormalized) Duh-Haymet plot for the LJ case. In contrast to the non-renormalized case there is a characteristic turn on the positive side which correlates roughly with the first maximum of the PDF, i.e. with the onset of the repulsive core. The curves continue on the left towards the limiting value at close contact; the attractive region is located around the origin. The bridge function shows apparently a bifurcation into two different functional dependencies. However, even in the attractive region (panel C of Figure 9) the bridge function can obviously not be fully described by a single function of t^*, although a

modified Verlet fit (which means a single function ansatz) appears to work well[78]. The most remarkable observation is, however, that all curves almost coincide in the attractive regions, independent of the thermodynamic state.

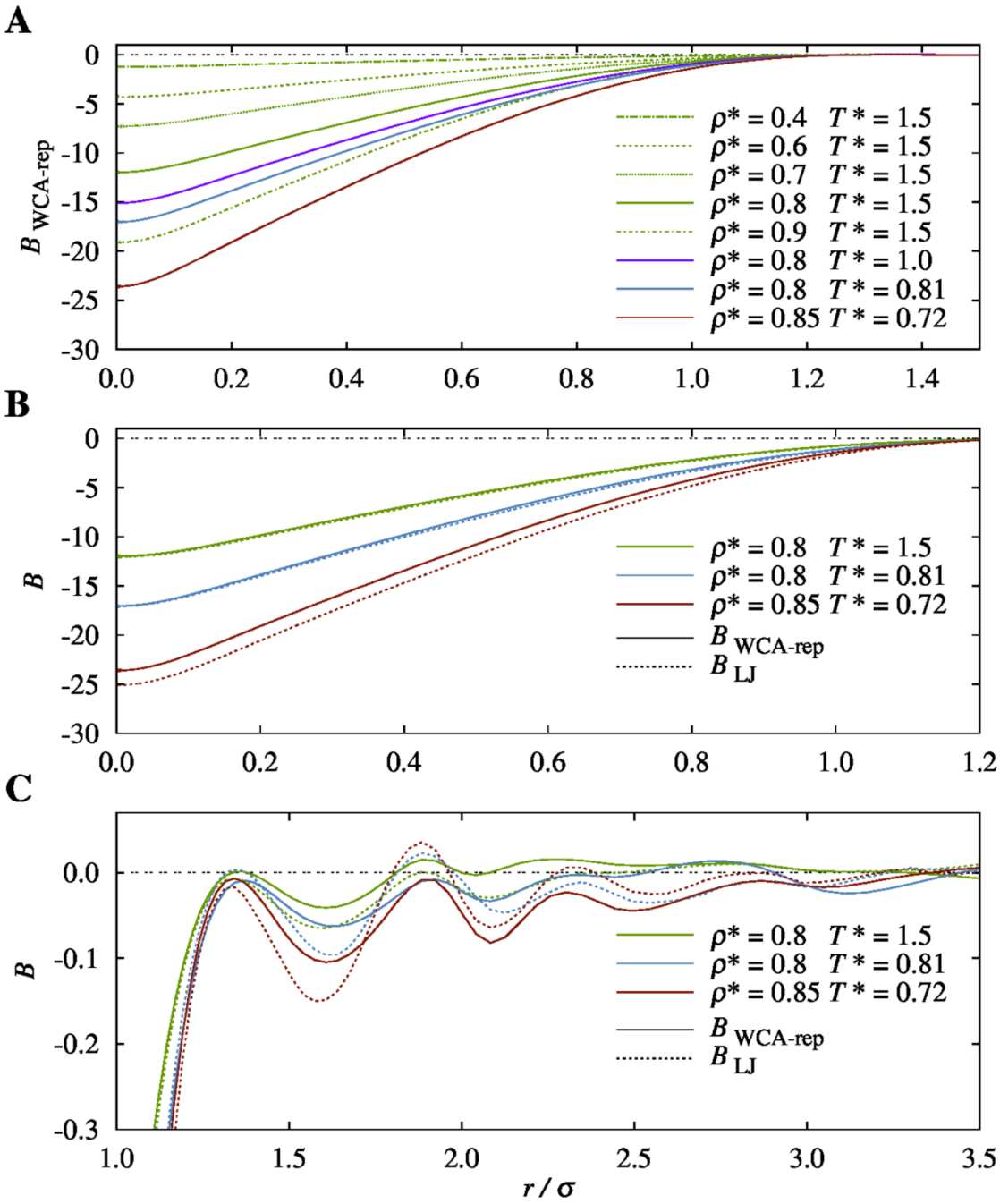


Figure 8 Bridge data from iterative inversion as a function of pair distance for various states represented by different line styles as specified in the panels; **A** for the repulsive WCA fluid, **B**,**C** comparison of repulsive WCA and full LJ fluids.

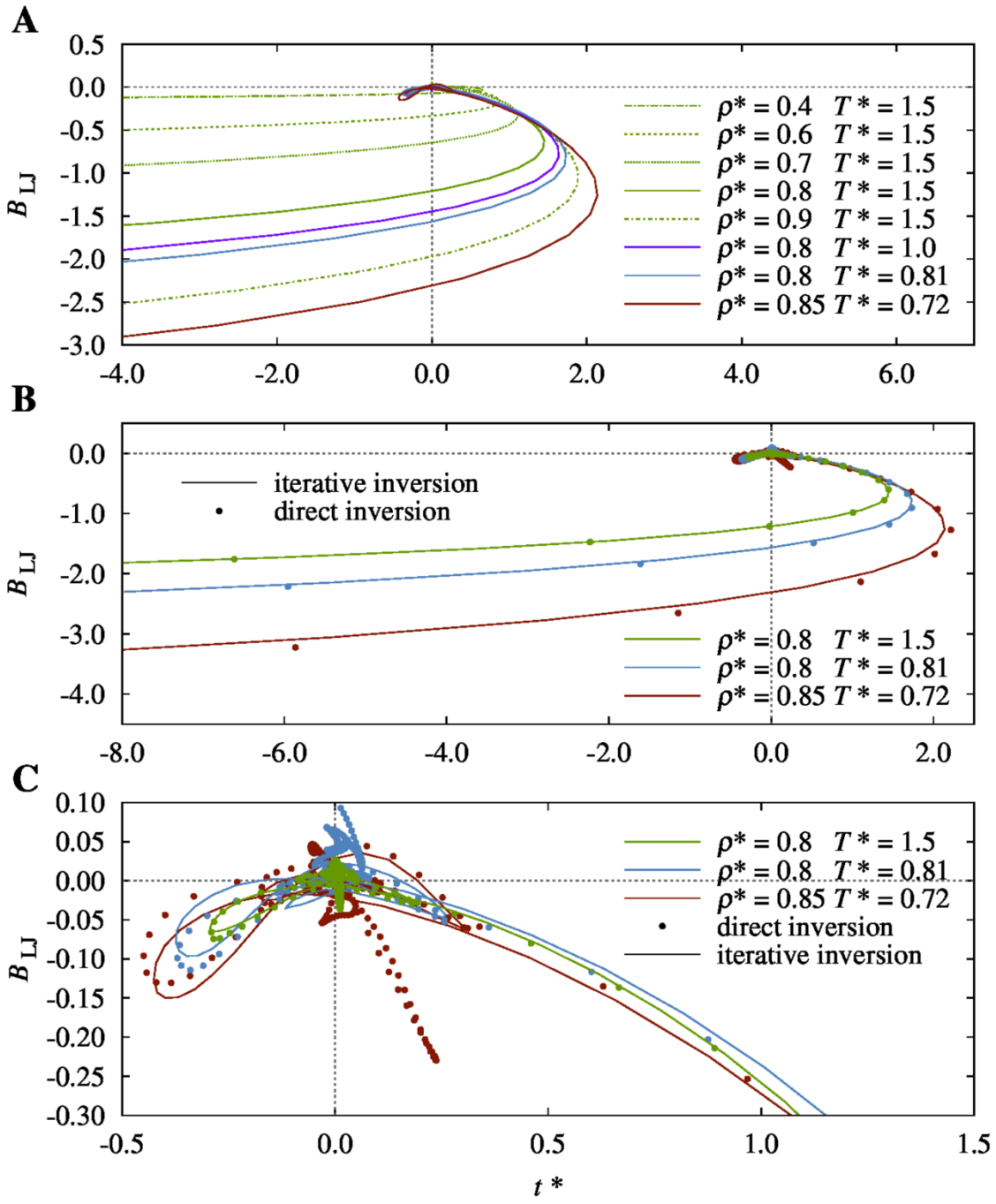


Figure 9. Bridge data for the LJ fluid from different inversion methods as a function of the renormalized ICF $t^*=t-\beta u$ for various states represented by different line styles as specified in the panels; **A**, **B**, **C** represent different plot ranges.

An analysis of the repulsive WCA case in comparison sheds more light on this situation. The numerical inversion results are shown in Figure 10. It turns out that all curves again almost coincide in the attractive potential region. The situation of Ref. [78] here corresponds to the transition between the solid and the dashed lines in the upper branch in panel B. The underlying assumption for a valid coupling parameter integration over a specific functional form of $B(t^*)$ is that this form is independent of the potential details, in this case for both the WCA-

rep and the full LJ cases. Figure 10 shows that this is indeed the case in the attractive, upper branch region. This is supported by the superimposed curve (dotted lines) of the VM closure parametrization presented in [78] which led to a successful free energy prediction. The VM bridge model practically coincides with both, the WCA-rep and the LJ bridges as a function of the respective ICFs in the attractive region. The WCA-rep bridge data obtained in this work therefore in retrospect validate the approach chosen earlier. The correspondence and the associated partial universality, however, break down in the repulsive core region. Implications for free energy functionals including bridge functions for the general case covering the full potential are the topic of ongoing work.

5.2.4 Concluding remarks

We have shown that explicit PMF simulations of constrained pairs yield reliable bridge function data inside the repulsive potential regions of the LJ and the repulsive WCA fluids, irrespective of the inversion approach (based on real or reciprocal space formalisms). Due to the nonlocality and nonlinearity of the OZ equation in conjunction with the closure relation it is very difficult to control the impact of statistical noise and other simulation features on the resulting bridge data. Even very accurate simulation data as obtained in this work result in small differences in the final target values due to practically uncontrollable propagation of error. For simple first order thermodynamic quantities such as the internal energy this is apparently unproblematic, whereas it remains to be seen if this is also the case for the free energy.

Improvement could possibly be achieved by smoothly interpolating the PMF approach to B in the core region with the direct inversion method applicable in regions where B is not superseded by the potential. We have not attempted such a method, nor have we played with smoothing techniques which all will yield further different results without a means to objectively measure the “correctness”. Another source of concern is the use of a directly inverted ICF (which would not be sufficient to obtain B in the core directly) in the context of a PMF-based bridge definition, an inconsistency which has also been noted earlier[33]. While there is reason to trust such an approach due to the insensitivity of the ICF to the core bridge function, many important questions arise. It has for instance to be clarified why correlation functions obtained by solving OZ/closure equations with adequate bridge data do not respond to the latter in the core region while at the same time this very core region has substantial impact on the free energy[112,8]. This problem together with attempts to understand the partial universality of the bridge functions in the attractive potential region and the discrepancies in the repulsive part are important areas for future research.

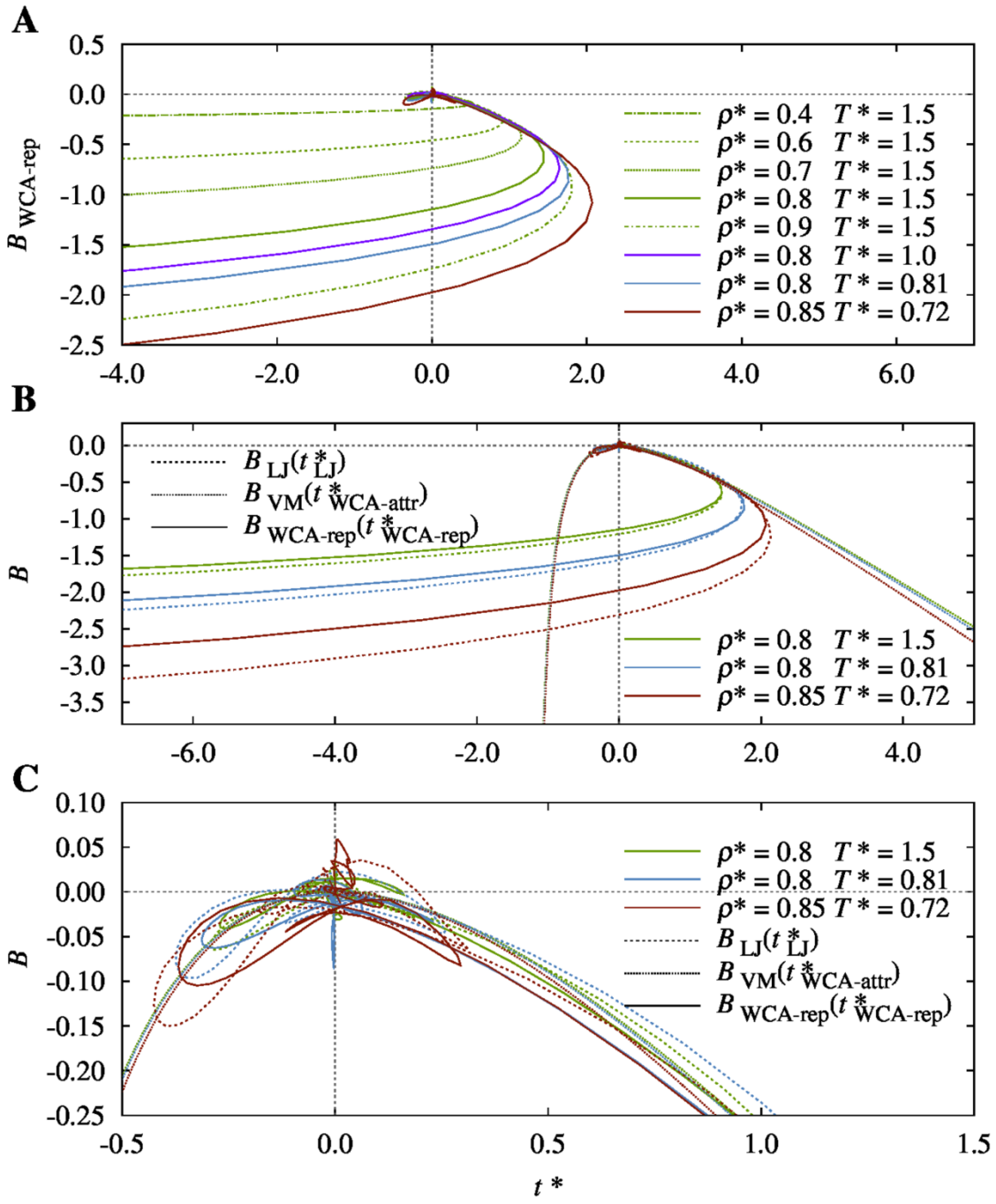


Figure 10 Bridge data for the repulsive WCA and for the LJ fluids from iterative inversion as a function of the renormalized ICF $t^*=t-\beta u$ for various states represented by different line styles as specified in the panels; **A** pure WCA-rep results for all states, **B**, **C** comparison with LJ data and with a partially (long range) renormalized bridge model described by the modified Verlet closure VM, parametrized corresponding to the free energy extraction approach of Ref [78].

6 Bridge function and RISM: Thermodynamics

As described in chapter 4.5 the thermodynamic properties in RISM calculations depend on the bridge function. In chapter 5 the exact bridge functions of some model fluids were given. The central goal of this chapter is to combine these facts to improve RISM calculations.

As pointed out by Kast[78] a bridge function has to be defined in dependence of either $t^* = t - \beta u$, h or $c^* = +\beta u$ or certain combinations thereof to get rid of the path dependence of the excess chemical potential as discussed in chapter 4.5.. If the bridge function is given as a function of the renormatized indirect correlation function the excess chemical potential in 1D-RISM is given by

$$\mu[B(t^*)] = \mu^{\mathrm{HNC}} + \frac{4\pi\rho}{\beta}\int r^2\left(h(r) - \int_0^{t^*} \mathrm{d}t^*(h(t^*;r)+1)\right)\mathrm{d}r\,. \tag{4.40}$$

The 3D-RISM functional is analog (without the $4\pi r^2$ term and integration over **r** instead of r). If the bridge function is given as a function of the total correlation function the formula for the excess chemical potential is[113]

$$\mu[h(B)] = \mu^{\mathrm{HNC}} + \frac{4\pi\rho}{\beta}\int r^2 \int_0^{B} g(B)\mathrm{d}B\,\mathrm{d}r\,. \tag{4.42}$$

In both cases the bridge function B and the "partner function" (t^* or g) depend on the separation r. Hence the equations (4.40) and (4.42) may be interpreted as a parametric function for every separation r.

In this chapter possible properties implied by this formalism and possible applications will be discussed. In the first part the equations (4.40) and (4.42) are discussed in relation to the zero separation theorem[6,7,8,9]. Then the validity of equation (4.40) is proven using the example of an alchemical transition process. Finally the equations (4.40) and (4.42) are used for absolute free energy calculation attempts. This chapter is exclusively about monoatomic simple fluids. In this case the reference interaction site model is exactly the Ornstein-Zernike equation. The term RISM will be used (instead of Ornstein-Zernike) for the sake of consistency with the other chapters.

6.1 Zero separation theorem

For a hard sphere fluid it is know that the excess chemical potential is directly related to the bridge function and the indirect correlation function at zero separation[8]

$$\beta\mu^{\text{ex}} = B(r=0) + t(r=0)\,. \tag{6.1}$$

Lee and Shing demonstrated that there are similar relations for all other fluids. But in this case they also depend on the interaction potential of the fluid[8]. Hence when using equation (4.40) or (4.42) for the calculation of excess chemical potentials a major contribution of the bridge function to the excess chemical potential is expected at low or zero separation.

6.2 Helmholtz energies

The equations (4.40) and (4.42) only apply to the excess chemical potential. In this chapter it will be demonstrated that a generalization to the Helmholtz energies is not straightforward. The Helmholtz energy can be calculated from 1D-RISM by

$$A^{\text{HNC}} = \frac{2\pi\rho}{\beta}\int r^2\left(\frac{1}{2}h^2(r) - c(r)\right)\mathrm{d}r + \frac{1}{4\pi^2\beta\rho}\int k^2\left(\rho\hat{c}(k) + \log[1-\rho\hat{c}(k)]\right)\mathrm{d}k\,. \tag{4.31}$$

6.3 Alchemical transition free energies

S. Kast[78] used a path integration to prove that path independence may be enforced, when the renormalized indirect correlation function is used to describe the bridge function. In his paper the Helmholtz energy of the LJ fluid was calculated using the WCA partitioning scheme. The Helmholtz energy of the repulsive core was approximated by the Helmholtz energy of a hard sphere fluid using the Carnahan-Starling[114] equation of state. Then the Helmholtz energy was calculated for the transition of the repulsive core to the LJ fluid. This was modeled using a coupling parameter path along which the attractive part of the LJ fluid was gradually switched on. The bridge function was chosen as a modified Verlet closure

$$B(t^*) = \frac{-\zeta}{2}t^{*2}\,\frac{1}{1+0.8t^*} \tag{4.18}$$

with the adjustable parameter ζ and

$$t^* = t - \beta u_{\text{WCA-atr}}\,. \tag{6.2}$$

The excess chemical potential of this transition is calculated using equation (4.40). The results are compared to MD simulations. The excess chemical potential change is given by

$$\Delta\mu[B(t^*)] = \mu[h_1, c_1, B_1] - \mu[h_0, c_0, B_0]\,. \tag{6.3}$$

Here the state "1" denotes the final state where all atoms in the system interact by using the LJ potential ($u = u_{\mathrm{LJ}}$). State "0" is the initial state where one WCA-rep particle is placed in a LJ solvent. An insertion of equation (4.40) in equation (6.3) leads to

$$\begin{aligned}\Delta\mu[B(t^*)] = {} & \mu^{\mathrm{HNC}}[h_1, c_1, B_1] - \mu^{\mathrm{HNC}}[h_0, c_0, B_0] \\ & + \frac{4\pi\rho}{\beta}\int r^2\, h_1(r)\mathrm{d}r + \frac{4\pi\rho}{\beta}\int r^2 \left(h_1(r) - \int_0^{t_1^*} \mathrm{d}t^*(h_1(t_1^*; r) + 1) \right)\mathrm{d}r \\ & - \frac{4\pi\rho}{\beta}\int r^2\, h_0(r)\mathrm{d}r - \frac{4\pi\rho}{\beta}\int r^2 \left(h_0(r) - \int_0^{t_0^*} \mathrm{d}t^*(h_0(t_0^*; r) + 1) \right)\mathrm{d}r.\end{aligned} \tag{6.4}$$

The bridge function is described by a modified Verlet bridge function with the same coefficients for the LJ and WCA-rep fluid. One can separate the λ dependent parts of the integration by definition of two auxiliary functions S_{VM} and T_{VM}

$$\begin{aligned} S_{\mathrm{VM}}^i &= \frac{1}{2}h_i(r) - \frac{1}{2}h_i(r)c_i(r) - c_i(r) + h_i(r), \\ T_{\mathrm{VM}}^i &= \int_0^{t_i^*(r)} \mathrm{d}t^*\, \exp[t^*(r) + B(t^*; r)]. \end{aligned} \tag{6.5}$$

In this case the modified Verlet closure (chapter 4.3.3) is used to replace the $h(r)$+1 term. Then $\Delta\mu[B(t^*)]$ is given by

$$\Delta\mu[B(t^*)] = \frac{4\pi\rho}{\beta}\int S_{\mathrm{VM}}^1 - S_{\mathrm{VM}}^0\mathrm{d}r + \frac{4\pi\rho}{\beta}\int T_{\mathrm{VM}}^1 - T_{\mathrm{VM}}^0\mathrm{d}r\,. \tag{6.6}$$

The differences h, h^2, c and hc terms of the S_{VM} functions are self-explanatory. The t^* integral T_{VM} of the LJ fluid is

$$\begin{aligned} t^* &= t_{\mathrm{LJ}} - \beta u_{\mathrm{WCA-atr}}, \\ T_{\mathrm{VM}}^1 &= \int_0^{t_{\mathrm{LJ}}^*(r)} \mathrm{d}t^* \exp[t_{\mathrm{LJ}} - \beta u_{\mathrm{LJ}} + B(t^*; r)], \\ T_{\mathrm{VM}}^1 &= \int_0^{t_{\mathrm{LJ}}^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_{\mathrm{WCA\text{-}rep}} + B(t^*; r)], \\ .T_{\mathrm{VM}}^1 &= \int_0^{t_1^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_0 + B(t^*; r)]. \end{aligned} \tag{6.7}$$

The t^* integral T^0_{VM} of the LJ WCA-rep fluid is

$$\begin{aligned} t^* &= t_{\mathrm{WCA\text{-}rep}}, \\ T^0_{\mathrm{VM}} &= \int_0^{t^*_{\mathrm{WCA\text{-}rep}}(r)} \mathrm{d}t^* \exp[t_{\mathrm{WCA\text{-}rep}} - \beta u_{\mathrm{WCA\text{-}rep}} + B(t^*;r)], \\ T^0_{\mathrm{VM}} &= \int_0^{t^*_{\mathrm{WCA\text{-}rep}}(r)} \mathrm{d}t^* \exp[t^* - \beta u_{\mathrm{WCA\text{-}rep}} + B(t^*;r)], \\ T^0_{\mathrm{VM}} &= \int_0^{t_0^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_0 + B(t^*;r)]. \end{aligned} \tag{6.8}$$

The difference $T^1_{\mathrm{VM}} - T^0_{\mathrm{VM}}$ is given by

$$\begin{aligned} T^1_{\mathrm{VM}} - T^0_{\mathrm{VM}} &= \int_0^{t_1^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_0 + B(t^*;r)] - \int_0^{t_0^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_0 + B(t^*;r)] \\ &= \int_{t_0^*(r)}^{t_1^*(r)} \mathrm{d}t^* \exp[t^* - \beta u_0 + B(t^*;r)]. \end{aligned} \tag{6.9}$$

Hence equation (6.6) turns into

$$\begin{aligned} \Delta\mu[B(t^*)] = \frac{4\pi\rho}{\beta} \int r^2 \Bigg(& \frac{1}{2}(h_1^2(r) - h_0^2(r)) - \frac{1}{2}(h_1(r)c_1(r) - h_0(r)c_0(r)) + (h_1(r) - h_0(r)) \\ & - (c_1(r) - c_0(r)) - \int_{t_0^*(r)}^{t_1^*(r)} \mathrm{d}t^* \exp[-\beta u_0(r) + t^* + B(t^*;r)] \Bigg) \mathrm{d}r. \end{aligned} \tag{6.10}$$

In the following chapters the excess chemical potential will be evaluated using different closures, bridge functions and excess chemical potential functionals. To allow for an easy identification how the excess chemical potential is calculated here, the following conventions will be used.

- F describes the excess chemical potential functional. F[HNC] means equation (4.30), F[PSE] means equation (4.41), $F[B(t^*)]$ means equation (4.40) and $F[g(B)]$ means equation (4.42).
- B defines the closure or the source of the bridge function used in the RISM calculation. B[HNC] is the bridge function from the HNC closure ($B(r) = 0$). B[MD] means the RISM calculations were performed with a bridge function extracted from MD simulations.
- χ is the solvent susceptibility function used in the RISM calculation. This is only important for $\mathrm{RISM_{uv}}$ calculations.

As an example the excess chemical potential calculated with the HNC functional of a $RISM_{uv}$ calculation in the LJ solvent with a bridge function from MD simulations is $\mu(F[\mathrm{HNC}], B[\mathrm{MD}], \chi[\mathrm{LJ}])$.

In RISM the free energy is calculated from an integration of spatial coordinates r. For example see equations (4.30), (4.40) or (6.10). The integrand is called excess chemical potential density ρ_μ. As an example the excess chemical potential density of the HNC functional is

$$\rho_\mu^{\mathrm{HNC}} = \frac{1}{2}h^2(r) - \frac{1}{2}h(r)c(r) - c(r). \tag{6.11}$$

6.3.1 Computational details

The free energy simulations were performed using LAMMPS[105]. The force field parameters were set to $\varepsilon_{\mathrm{LJ}}$ = 0.0709 kcal/mol and σ_{LJ} = 2.79 Å for both the LJ and the WCA-rep fluid. In the case of the LJ fluid a potential cutoff at 10 Å was used. The thermodynamic integration was performed along a linear path. The path was sampled at 11 points between λ = 0.01 and λ = 0.99 (0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99). The partially coupled systems were calculated using cubic spline potentials. The potential was calculated on a linear grid of 10,000 points between 0.01 Å und 10 Å.

Table 3: States of the LJ and WCA-rep fluids used in the MD simulations calculated with equation (3.17) from the reduced temperature and density. The temperature was calculated with $\varepsilon_{\mathrm{LJ}}$ = 0.0709 kcal/mol as well as k_{B} = 0,001987 kcal/mol and the density was calculated with σ_{LJ} = 2.79 Å.

T^*	ρ^*	T / K	ρ / Å^3
1.5	0.4	53,52	0,01842
1.5	0.6	53,52	0,02763
1.5	0.7	53,52	0,03223
1.5	0.8	53,52	0,03684
1.5	0.9	53,52	0,04144
1.0	0.8	35,68	0,03684
0.81	0.8	28,90	0,03684
0.72	0.85	25,69	0,03914

Each simulation was equilibrated for 50 ps and then simulated for 500 ps starting from a random distribution of the particles. The time step was set to 1 fs. The simulations were performed in the NVT ensemble using a Nose-Hoover thermostat[106,107]. The density and tem-

perature was set to the combinations listed in table 3. All simulations were performed with 3000 particles. For the calculation of excess chemical potentials the interaction of one particle with every other particle was scaled. For the Helmholtz energy the potentials of all interactions were scaled. The errors of the simulated thermodynamic properties were estimated using blocking analysis[115].

1D-RISM calculations were done on a 512 point logarithmically spaced grid ranging from 0.0059 Å to 164.022 Å. The LJ parameters were set to ε_{LJ} = 0.079 kcal/mol and σ_{LJ}= 2.79 Å as in the MD simulations. No cutoff was applied to the LJ potential. Two 1D-RISM calculations were performed to calculate the excess chemical potential:

1. 1D-RISM$_{vv}$ for the LJ fluid as a representation of the final state,
2. 1D-RISM$_{uv}$ for one WCA-rep particle in the LJ fluid. This corresponds to the starting point of the calculation of excess chemical potential.

The 1D RISM calculations were performed two times. First with the HNC closure and second with the modified Verlet closure with the ζ parameters taken from S. Kast[78]. The RISM iterations were performed until changes of the direct correlation functions fell below 10^{-5} at all separations from one iteration to the next.

6.3.2 Results

Table 4 shows the excess chemical potentials of the transition of a WCA-rep participle to a LJ participle in the LJ fluid. The excess chemical potentials from MD simulations are in good agreement with the RISM calculations of both closures (HNC and modified Verlet), when the t^* integration is added. Figure 11 shows the excess chemical potential calculated with the HNC functional and corrected with the Verlet bridge functional. In panel **A** the Verlet closure was used in the RISM calculations. Using the conventions form page 50 the excess chemical potentials are $\mu(F[\text{HNC}], B[\text{VM}], \chi[\text{LJ}])$ in green and $\mu(F[B(t^*)], B[\text{VM}], \chi[\text{LJ}])$ in red. The susceptibility function χ[LJ] was calculated with the modified Verlet closure as well. In panel **B** the HNC closure was used for the RISM$_{uv}$ calculations and the RISM$_{uu}$ calculations (for the susceptibility function). Here the excess chemical potentials $\mu(F[\text{HNC}], B[\text{HNC}], \chi[\text{LJ}])$ in green and $\mu(F[B(t^*)], B[\text{HNC}], \chi[\text{LJ}])$ in red are shown. In both cases the bridge dependent t^* functional (red) with the Verlet bridge function improves the RISM calculations. Surprisingly the correction works better when paired with the HNC closure (panel **B**) than with the Verlet closure (panel **A**).

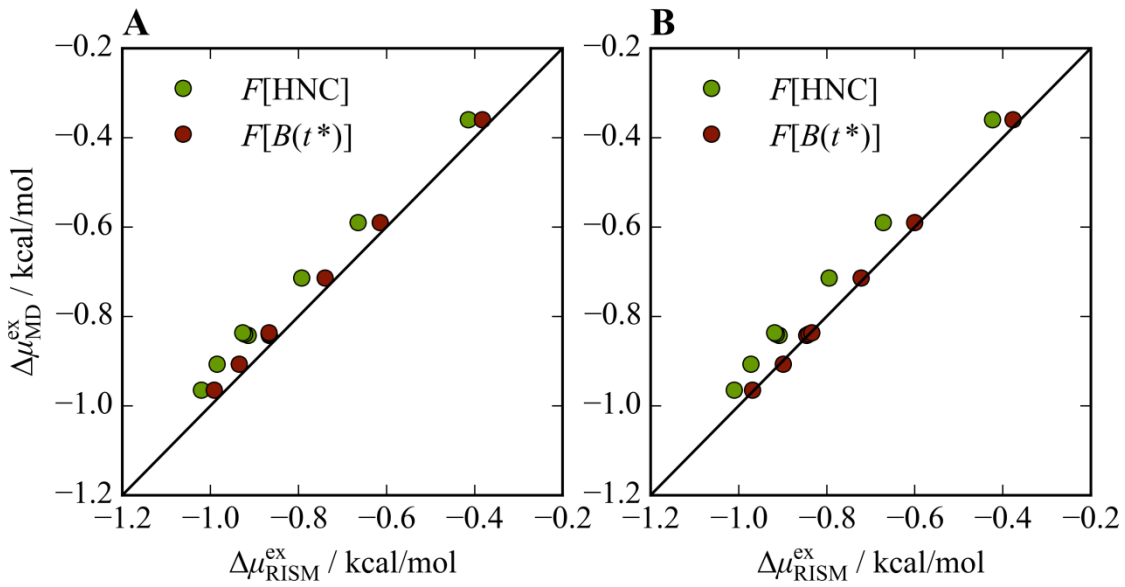


Figure 11: Excess chemical potential change for the alchemical transition of a WCA-rep particle to a LJ particle at different densities and temperatures. The green circles use the HNC functional of the excess chemical potential and the red circles use a correction of the excess chemical potential based on the t^* integral with simulated bridge functions. Panel **A** shows the results for 1D-RISM with the modified Verlet closure and the HNC functional (red) or t^* functional (green). Green: $\mu(F[\mathrm{HNC}], B[\mathrm{VM}], \chi[\mathrm{LJ}])$, red: $\mu(F[B(t^*)], B[\mathrm{VM}], \chi[\mathrm{LJ}])$. Panel **B** shows the results for 1D-RISM with the HNC closure. Green: $\mu(F[\mathrm{HNC}], B[\mathrm{HNC}], \chi[\mathrm{LJ}])$, red: $\mu(F[B(t^*)], B[\mathrm{HNC}], \chi[\mathrm{LJ}])$.

Figure 12 shows the excess chemical potential density ρ_μ at different states divided into the contributions of HNC functional and the Verlet bridge dependent functional. The contribution of the bridge dependent functional is small but it improves the accuracy of the calculations in all cases. The bridge functional peaks at a separation corresponding to the LJ and WCA σ parameter. This is the area where the LJ and WCA-rep potential are most different. The functional is zero close to zero separation. The bridge functions of the LJ fluid and the WCA-rep fluid are similar in this area. In table 4 the transition excess chemical potentials calculated with the modified Verlet closure and t^* integration is compared to the results from MD simulations and 1D-RISM/HNC. The analytical evolution of the t^* integral is an improvement over the HNC. In the present case, the modified Verlet closure is a good approximation for the bridge function and works well in combination with the t^* integration for the bride dependent functional.

6.4 Absolute excess chemical potentials

The bridge dependent functional can correct excess chemical potentials for the transition of a WCA-rep to a LJ particle. This chapter is about the calculation of absolute excess chemical potentials at the example of the LJ fluid. In the previous chapter the initial state “0” was the

WCA-rep fluid and the final state "1" was the LJ fluid. Hence initial and final states were similar and the calculations may benefit from error canceling. Now the initial state is the potential free homogeneous fluid. This is a more general but also more complicated case.

Table 4: Excess chemical potential difference $\Delta\mu^{ex} / \varepsilon_{LJ}$ of the transition of a WCA-rep particle to a LJ particle in the LJ fluid at different states. The values were calculated from MD simulations, 1D-RISM with the modified Verlet closure and t^* integration at the bridge dependent functional and 1D-RISM with the HNC closure.

T^*	ρ^*	MD	Verlet	HNC
1.5	0.4	-0.36 ± 0.01	-0.38	-0.42
1.5	0.6	-0.590 ± 0.006	-0.614	-0.672
1.5	0.7	-0.714 ± 0.005	-0.739	-0.794
1.5	0.8	-0.843 ± 0.004	-0.866	-0.908
1.5	0.9	-0.965 ± 0.002	-0.991	-1.010
1.0	0.8	-0.840 ± 0.004	-0.867	-0.915
0.81	0.8	-0.837 ± 0.004	-0.867	-0.918
0.72	0.85	-0.90 ± 0.06	-0.93	-0.97

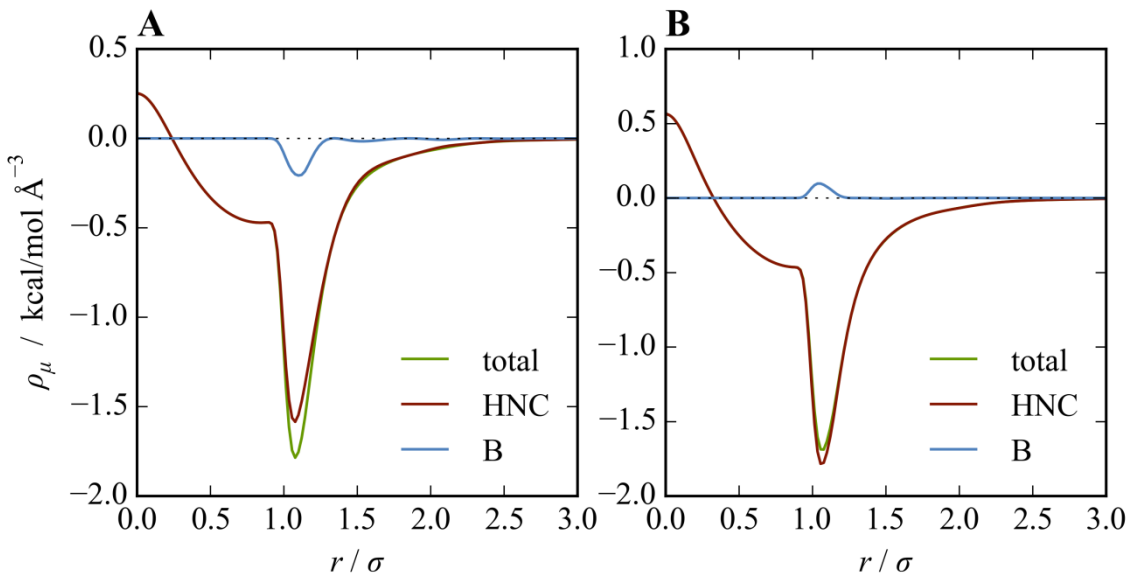


Figure 12: Excess chemical potential density $\rho\mu$ of the transition of a WCA-rep particle to a LJ particle in the LJ fluid. The state is $\rho^* = 0.8$ and $T^* = 1.0$. The 1D-RISM calculations were done with the modified Verlet closure. The free energy is given for the HNC functional $\mu(F[\text{HNC}],B[\text{VM}],\chi[\text{LJ}])$ (red), the bridge dependent functional using a t^* integration $\mu(F[B(t^*)\text{-HNC}],B[\text{VM}],\chi[\text{LJ}])$ (blue) and the sum of both $\mu(F[B(t^*)],B[\text{VM}],\chi[\text{LJ}])$ (green). The free energy density is shown for the usage in equations (4.30) and (4.40). Hence the special factor $4\pi r^2$ is not included.

6.4.1 Parametric bridge functions

The bridge function of the LJ fluid can be given as a parametric function of the renormalized indirect correlation function. In this case the excess chemical potential is given by equation (4.40). Since t^* is a parametric function of the separation r, all integration in (4.40) can be separated into an in-core contribution from zero separation to a border r_b and from this border to infinite separation. As described in chapter 4.4.1 and in figure 13 **A** the graph of $B[t^*]$ vs t^* cannot be described by a single function. However the attractive region (corresponding to separations bigger than the LJ σ parameter) can be approximated using a modified Verlet fit. The repulsive part can be modeled as a spline. Therefore the border r_b can be adjusted to the value where the Verlet function is replaced by the spline function.

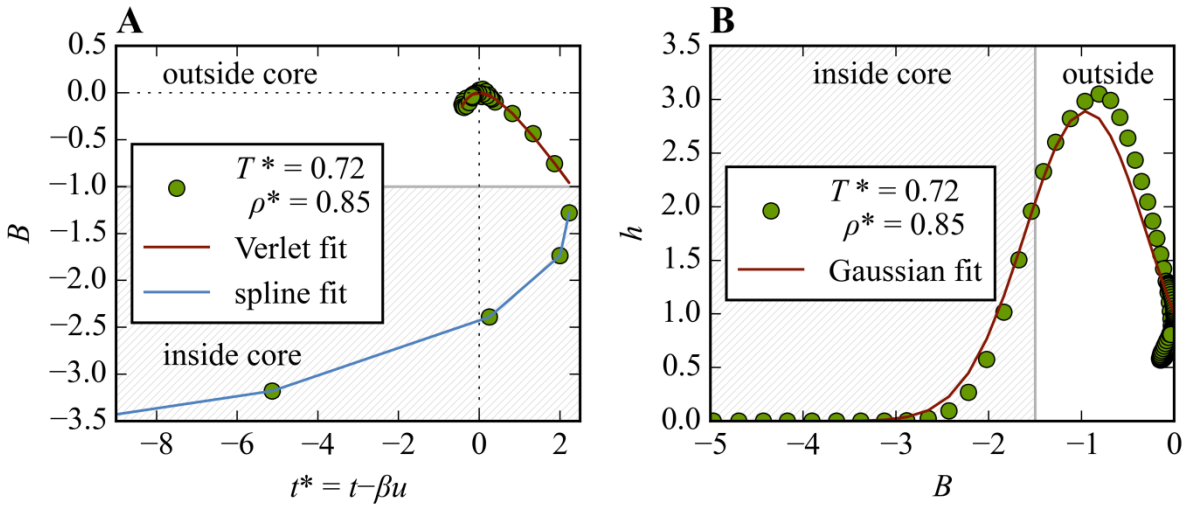


Figure 13: Comparison of different bridge dependent functionals. Panel **A** shows the t^* integration and panel **B** the $g(B)$ integration. In both panels a parametric representation of the bridge function and a correlation function is show. The parts of the parametric functions corresponding to separations smaller than 1.1 σ_{LJ} are labeled as inside core.

6.4.1.1 Computational details

The 1D-RISM calculations were performed on a logarithmically spaced grid of 512 points ranging from 0.00155 σ to 48.1709 σ. The bridge function was taken from literature[89,78], either in the form of fitted ζ parameters for the VM closure or directly as a function of the separation. The border r_c was set to the separation corresponding to the maximum of the bridge function. The inside core bridge function was interpolated to the calculation grid using splines of order one without smoothing. The bridge function closer to the core than known from literature, was set to the first known value. The bridge function beyond the border was set to the modified Verlet closure using the Verlet parametrization form [78]. The t^* integration of both branches was performed numerically.

6.4.1.2 Results

The parametric description of the bridge function as a spline and via the Verlet function is shown in figure 13 **A**. The Verlet fit describes the outside core region well. The spline function is a good representation of the bridge inside the core in the parametric description. However, when the parametric dependency is resolved, the resulting bridge functions are not in perfect agreement with the published data (MD). Therefore the bridge function is evaluated in the parametric for a given t^* and this value is plotted against the separation r corresponding to t^*. This bridge function can then be compared with the original simulation data. The parametric resolved bridge function shows a discontinuity between the spline and the Verlet representation at the border of the core. The spline representation is inaccurate at low separations between the known values sowing that linear splines do not have the correct interpolation properties after the parametric dependency is resolved. This is shown in figure 14 **A**, **B** and **C**.

The excess chemical potential calculated with t^* integration (equation (4.40)) and the LJEoS (see chapter 3.3.1) is shown in table 5. The t^* integration is an improvement over the HNC. The $g(B)$ integration is comparable to the HNC results and does not improve the accuracy of the excess chemical potentials. The excess chemical potential density contribution of the t^* integration is relatively small inside the core. At the border the free energy density has an artificial peak as seen in panel **D** of figure 14.

Table 5: Excess chemical potential $\mu^{ex} / \varepsilon_{LJ}$ of the LJ fluid. The calculations were performed with 1D-RISM and the modified Verlet closure using the t^* and $g(B)$ integration and with 1D-RISM using the HNC closure. The 1D-RISM results are compared to the reference values from the LJEoS. The RMSE to the reference is given.

State	T^*	ρ^*	referecne	t^* integration	$g(B)$ integration	HNC
1	1.5	0.4	-1.94	-2.12	0.29	-1.79
2	1.5	0.6	-1.63	-1.89	-3.12	-0.77
3	1.5	0.7	-0.6	-0.66	-7.27	0.96
4	1.5	0.8	1.43	4.9	-1.61	3.97
5	1.5	0.9	4.87	9.53	-7.67	8.73
6	1.0	0.8	-2.27	-0.5	-5.67	-0.05
7	0.81	0.8	-3.86	-1.26	-8.55	-1.78
8	0.72	0.85	-3.93	1.2	-8.19	-1.38
			RMSE	2.96	8.35	7.92

An alternative to the t^* integration is the $g(B)$ integration as described in equation (4.42). A plot of the total correlation function as a function of the bridge function is panel **B** in figure 13. This functional is well behaved inside the core, but the oscillations of the total correlation function make the description of bigger separations complicated. However, a numerical evaluation of this integral may be performed to the separation where the oscillations of the total correlation function begin. The excess chemical potentials of the LJ fluid calculated this way are listed in the $g(B)$ column of table 5. The results show little correlation with the LJEoS. However, at large separation, the excess chemical potential density corresponds to the density of the t^* integration. This is an indication that the mathematical and physical basics described by Kast and Kloss[15] are correct in contrast to what is expected from the missing agreement with the LJEoS. The numerical integration of the parametric functions seems to be the problem. However, the contribution of the bridge function inside the core may be underestimated in the $g(B)$ integration.

Finally it can be said that the calculation of the excess chemical potential of the LJ fluid with the t^* integration was an improvement over the HNC. However, the resulting errors remain substantial. The $g(B)$ integration has a similar performance when compared to the HNC. Both bridge dependent functionals are numerically sensitive, but the underlying fits are well behaved.

From the zero separation theorem one has to expect that the value of the bridge function at zero separation is important for the calculation of excess chemical potentials. This is not supported by the excess chemical potential densities. The HNC functional density is much bigger than the density of the bridge dependent functional. The bridge function does not appear in the HNC functional. It is not directly influenced by the value of the bridge function. The indirect contributions of the bridge function to the HNC functional will be discussed in chapter 0.

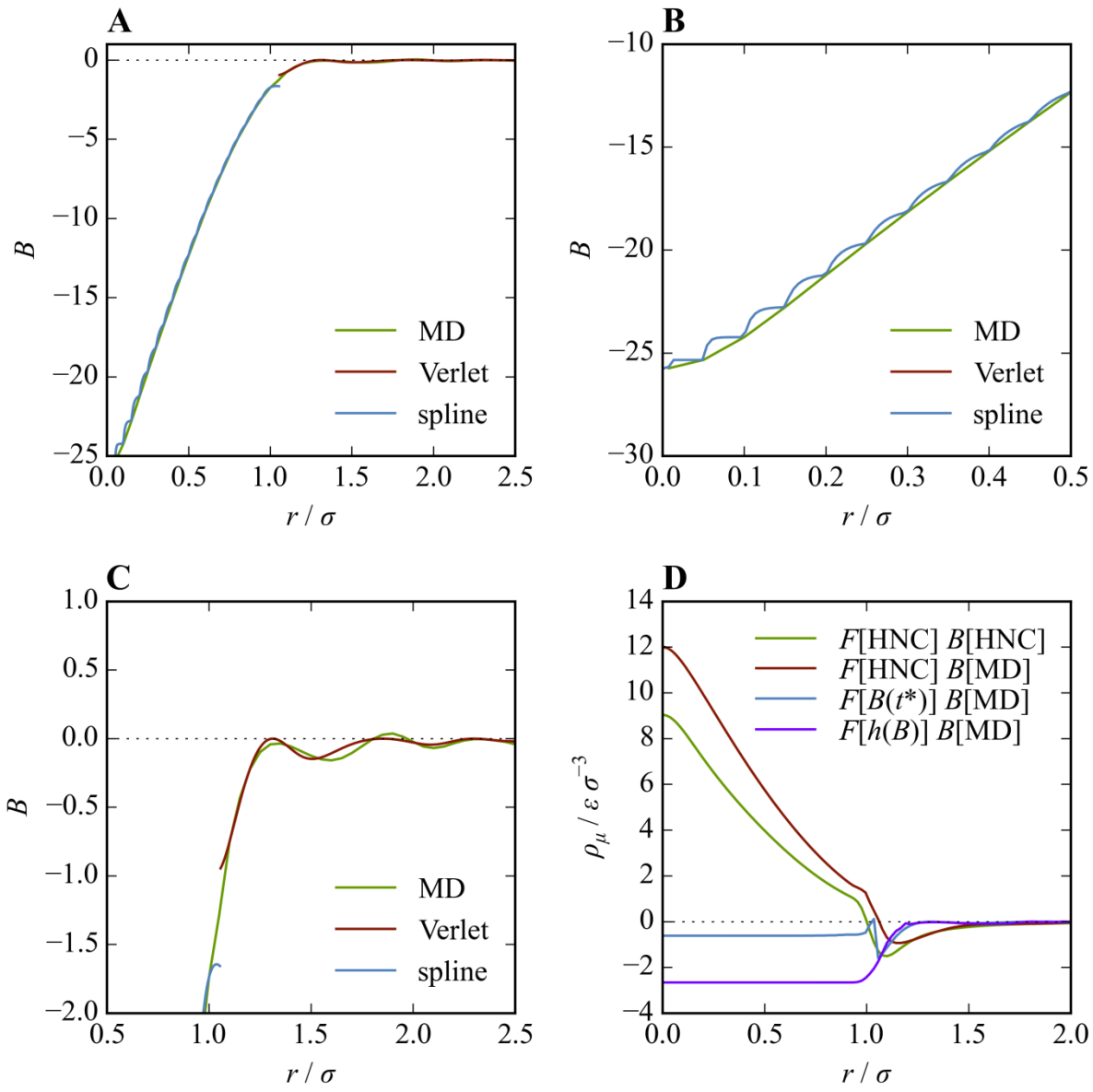


Figure 14: Panel **A**, **B** and **C** show the bridge function of the LJ fluid calculated from the modified Verlet closure and the spline in comparison to the simulated bridge function. Panel **D** shows the excess chemical potential density of the LJ fluid calculated using different bridge functions B and excess chemical potential functionals F. The free energy density is shown for the usage in equations (4.30), (4.40) and (4.42). Hence the factor $4\pi r^2$ is not included.

6.4.2 Relation of the excess chemical potential and the Helmholtz energy

S. Kast proposed[113] that the Helmholtz energy can be corrected in a similar way like the excess chemical potential. The equations (4.40) and (4.42) show that the excess chemical potential can be calculated by

$$\mu^{\text{ex}} = \mu^{\text{HNC}}[h_B, c_B] + \Delta\mu^B. \tag{6.12}$$

For the Helmhotz energy a similar relation should exist

$$A = A^{\text{HNC}}[h_B, c_B] + \Delta A^B = A^{\text{HNC}}[h_B, c_B] + \frac{1}{2}\Delta\mu^B. \tag{6.13}$$

The reason for this is that the λ integral in equation (4.34) is common in the expressions of ΔA^B and $\Delta\mu^B$

$$\Delta A^B = \frac{2\pi\rho}{\beta}\int r^2 \int (h_B + 1)\frac{\partial B}{\partial\lambda}\,\mathrm{d}\lambda\mathrm{d}r = +\frac{1}{2}\Delta\mu^B. \tag{6.14}$$

If this assumption is true, there has to be a bridge function that fulfills the relation

$$2(A - A^{\text{HNC}}[h_B, c_B]) = \mu^{\text{HNC}}[h_B, c_B] - \mu. \tag{6.15}$$

To test this assumption, a genetic algorithm was used to find a bridge function that fulfills equation (6.15). The bridge form Llano-Restrepo and Chapman[89] was used as a starting point of the optimization. The RISM calculations were performed as described above with a reduced temperature of 0.72 and a reduced density of 0.80. Within each iteration of the optimization the bridge function was "mutated" at random position and the functions with a higher "fitness" than the starting bridge functions were "recombined". The "fittest" bridge function of all trials in one iteration was used as a starting point for the next iteration. The "fitness" was measured by the squared difference between the right hand side and the left hand side of equation (6.15). As "mutations" ten values of the bridge function between the first and 400$^{\text{th}}$ (of a total of 512 grid points) grid points were randomly chosen. A Gaussian distributed perturbation with a standard deviation of 0.05 was added at these positions of the bridge function. The "recombination" step was done by the formation of random pairs between the bridge functions. The arithmetic average of the bridge functions was formed. The number of pairs was equal to the number of bridge functions selected for the recombination. A total of 530 iterations were performed.

The right hand side and left hand side of equation (6.15) are shown in table 6. The results show that there is a bridge function that fulfills the condition of equation (6.15). However,

figure 15 shows that this bridge function and the resulting pair distribution functions are substantially different to the results from the MD simulation. Therefore it is likely that the assumption of equation (6.15) has to be discarded. Though the formula of ΔA^B and $\Delta\mu^B$ are similar, the scaling of h_B is different when only on particle is introduced (μ) and when all particles are introduced (A) leading to different numerical values of the integral. However as shown in the next chapter the energy evaluation can be very sensitive against changes of the bridge function. This and the relatively simple genetic algorithm used here may have influenced the results.

Table 6: Comparison of the expected values of the bridge dependent functionals for the excess chemical potential and the Helmholtz energy. The bridge function was iteratively modified to match the result of both expressions. The corresponding bridge function and pair distribution function are shown in figure 15.

	$2(A - A^{\mathrm{HNC}}[h_{\mathrm{B}},c_{\mathrm{B}}]) / \varepsilon_{\mathrm{LJ}}$	$\mu^{\mathrm{HNC}}[h_{\mathrm{B}},c_{\mathrm{B}}] - \mu / \varepsilon_{\mathrm{LJ}}$
after 530 iterations	14.32	15.50
reference bridge function	13.90	22.23

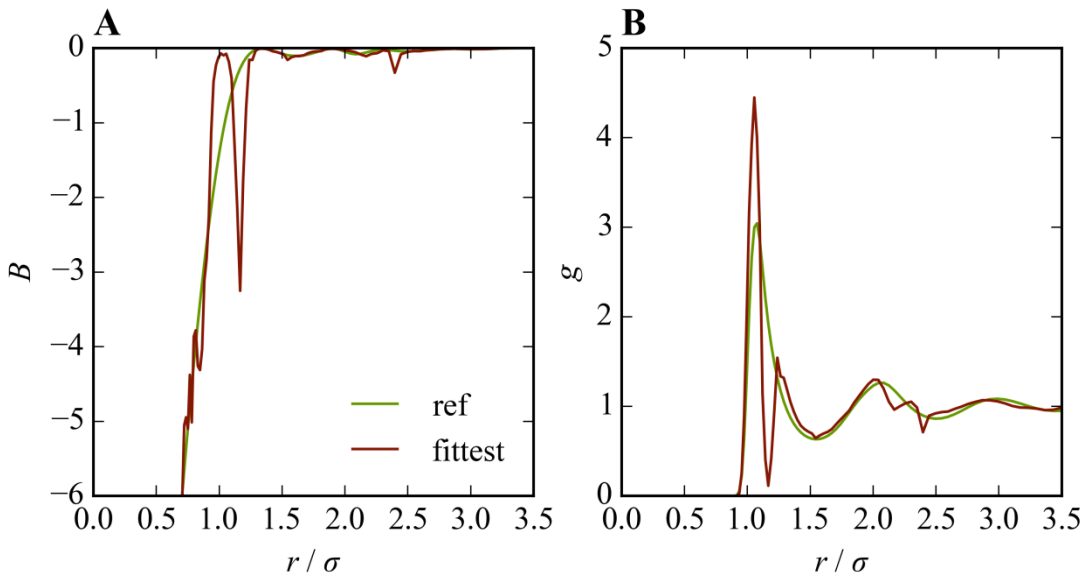


Figure 15: Panel **A**: Comparison of the bridge function from MD simulations (green) and the bridge function optimized for consistency of the excess chemical potential and Helmholtz energy (red) . Panel **B:** The resulting pair distribution functions.

The comparison of the bridge functions in figure 15 (panel **A**) shows a prominent peak at 1.2 Å. In this point the numerical stability of the RISM calculation seems to be surprisingly high. This is, together with a second smaller peak at 2.5 Å (and the region inside the core), the only

region where significant changes of the bridge function are tolerated. Consequently these regions are drastically changed to reach the optimization target.

6.4.3 Test of the zero separation theorem

As one can see in 6.4.1.2 the excess chemical potential densities of the bridge dependent functionals do not contribute significantly to the excess chemical potential at zero separation. With the zero separation theorem it is possible to estimate whether these functionals must have an influence or not. When the HNC functional describes the excess chemical potential at zero separation without a contribution of the bridge dependent functional, the value of the bridge function close to zero has to have a huge impact on the HNC functional due to the zero separation theorem. To be able to study this, perturbations were added to the bridge at close separations and the HNC functional was evaluated.

A normally distributed noise with the standard deviation σ was added to parts (between r_1 and r_2) of the bridge function of Llano-Restrepo and Chapman[89]. See figure 16 **A** as an example. The resulting bridge function was smoothed using approximate splines with a smoothing coefficient[116,117] α (see figure 16). The modified bridge was used in 1D-RISM calculations as described before. As seen in figure 16 **B** the perturbation has little impact on the HNC functional when it is located inside the core only. However, a perturbation in the area beyond the σ_{LJ} parameter influences the HNC functional. As expected, the excess chemical potential may be raised or lowered by the perturbation.

In the closure relation the bridge competes with the pair potential. At lower separations than σ_{LJ} the LJ potential energy is much higher than the bridge which renders the bridge function insignificant. As an example the bridge of the LJ fluid at T^*=0.72 and ρ^* = 0.80 is about 30 while $\beta u(r = \sigma_{LJ}/2)$ is about 20.000 ε_{LJ}. This means that the closure relation does not transport local information about the birdge funciton at close seperation to the correlation function. Therefore the bridge function at zero distance does not affect the HNC functional of the excess chemical potential. However, since RISM involves a convolution, the correlation functions at zero seperation are affected by the bridge in general. The HNC functional is sensitive towards perturbations from about a distance of σ_{LJ} onwards. In case of an absolute pertubation (such as in this study) the response increases with the seperation. Here the absolute values of the bridge function and the interaction potentials are small so the perturbation has a higher relative effect. This is shown in table 7. Since the HNC functional is unaffected by the bridge function at zero separation, this influence at zero sepertion has to come from the bridge dependent functional. This is a central statement of the zero seperation

theorem. Here exists a central issue of the $g(B)$ integral. Inside the core, the pair distribution function g rapidly approaches zero. Consequently an integration of g over B will not include the value of B at small separation because g is zero at these points.

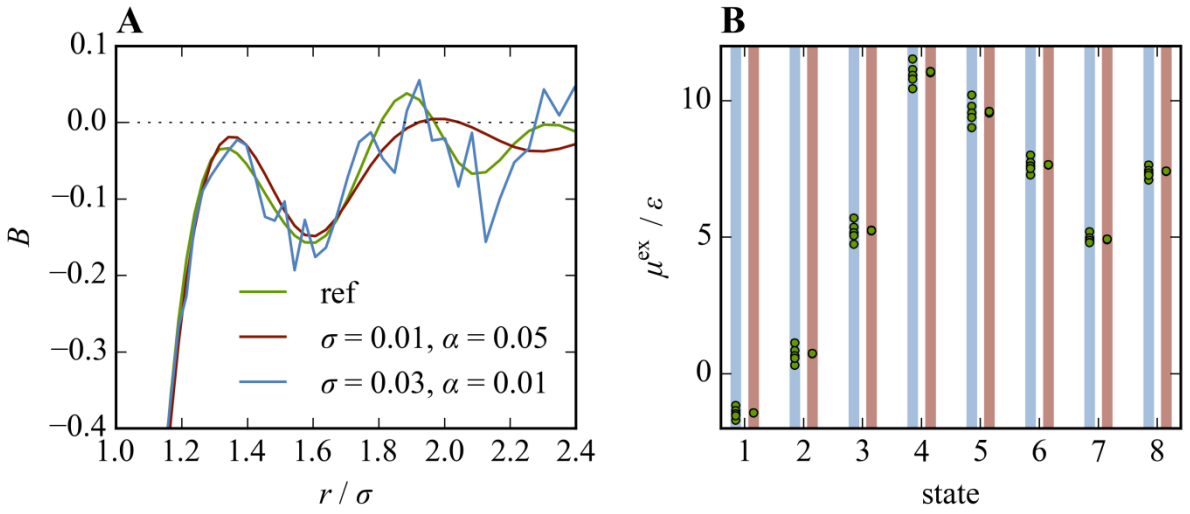


Figure 16: Panel **A:** Randomly perturbed bridge functions between 1.0 σ_{LJ} and 2.5 σ_{LJ}. Panel **B:** resulting excess chemical potentials. The bars show the parameters of the perturbation and the **green** circles are excess chemical potentials calculated with the HNC functional of the bridge perturbation labeled blue or red. The bar colors are: **Red:** Perturbation with the standard deviation of σ = 0.0036 and a smoothing of α = 0.01 between 0.036 σ_{LJ} and σ_{LJ}. **Blue:** Perturbation between 2.0 σ_{LJ} and 3.6 σ_{LJ} with the same σ_{LJ} and α. The states (1-8) correspond to table 5. From the free energies of state 5 a value of 15 ε^{-1} was subtracted from all excess chemical potentials to match the plotted range.

As a conclusion both discussed bridge dependent functionals (the t^* integration and the $g(B)$ integration) have different drawbacks. The t^* integration requires integration over the renormalized indirect correlation function. Since this value approaches minus infinity at close separation and the bridge function has a finite but non-zero value at zero separation, the integration is prone to divergence. Also there is the transition from the inside core area to the outside core area where a discontinuity of the integration is introduced. The $g(B)$ integration does not fulfill the zero separation theorem and ignores the out of core area. To provide a purely physically based approach towards the calculation of correct excess chemical potentials of the LJ fluid, the fundamentals of RISM are not well enough understood yet. As a consequence this work will now focus on how RISM can be improved using knowledge of the system itself or a similar reference system. Therefore a semi-empirical combination of RISM derived quantities with quantities from other sources like experimental data will be used.

In this case the partial molar volume (4.43) is an obvious choice. It can be calculated by integration over the direct correlation function. The direct correlation function (figure 4) has a very basic similarity with the free energy density (figure 14). Most importantly, it is non-zero in the core and decays to zero with increasing separation. Hence a scaled and eventually shifted integral over the direct correlation function may be a good approximation of the free energy density. In the next chapter a semi-empirical correction based on the partial molar volume for EC-RISM is parameterized and discussed.

Table 7: Standard deviation of the HNC functional of the excess chemical potential when a randomly perturbed bridge function is used. The perturbation is a Gaussian noise with σ = 0.0036 σ_{LJ} and an approximate spline smoothing with α = 0.01 between r_1 and r_2. The states (1-8) correspond to table 5.

r_1	r_2	1	2	3	4	5	6	7	8
0.036 σ_{LJ}	σ_{LJ}	0,0011	0,0020	0,0028	0,0038	0,0054	0,0018	0,0012	0,0012
σ_{LJ}	2 σ_{LJ}	0,0147	0,0242	0,0305	0,0286	0,0495	0,0293	0,0185	0,0208
2σ_{LJ}	3.58 σ_{LJ}	0,0920	0,1376	0,1606	0,1833	0,2052	0,1222	0,0767	0,0932

7 Semi-empirical corrections of RISM free energies

As shown in the previous chapters, excess chemical potentials calculated by using HNC-type closures (like PSE-*k*) are inaccurate. A proper description of the bridge function contributions to the excess chemical potential is needed. Anyway at the moment there is no straightforward way to calculate this contribution even for simple fluids. Since the physical background is not yet completely understood, semi-empirical correction can be used to approximate the bridge function contributions to the free energies. In this chapter free energies of small organic molecules are calculated using EC-RISM. A semi-empirical correction for these energies is derived using experimental hydration free energies from the Minnesota solvation database[18].

7.1 Calculation of small molecule hydration free energies

The calculation of solvation free energies of small molecules is one of the challenges for theoretical chemistry nowadays. Many research groups have published about this and even contests like the SAMPL[118,119,120,121] challenges took place. This chapter gives an overview of physics based (semi-empirical) predictions of solvation free energies of small molecules. Then the focus is shifted to studies using the reference interaction site model.

An overview of the performance of the prediction of hydration free energies is given in the concluding paper to the SAMPL4 contest[121]. In the conclusions the authors say that “many methods are apparently converging on robust, predictive protocols with RMS errors under 1.5 kcal/mol”. Also they “find a relatively wide range of methods perform quite well on this test set, with RMS errors in the 1.2 kcal/mol range for several of the best performing methods”. The experimental solvation free energies were within 20 kcal/mol (as a comparison the range of solvation free energies of the Minnesota solvation database is about 100 kcal/mol). Here it is important to take note of the fact that this is a blind study. The RMSEs of the protocols will be lower for the training sets.

The free energy of solvation can be calculated by using molecular dynamics in combination with free energy simulation techniques. MD simulations usually ignore atom polarizability, but take the flexibility of the molecule into account and use explicit solvation. Shivakumar[122] *et al.* reported a mean unsigned error of 0.7 kcal/mol in a study of 239 neutral small molecules using the OPLS2.0 force filed. Fennell[123] *et al.* used MD simulations with a set of 504 small neutral molecules with the GAFF force field to compare implicit and explicate solvation mod-

els. With an implicit solvent model (GB) they achieved a RMSE of 2.8 kcal/mol and with explicit TIP3P water the RMSE decreased to 1.2 kcal/mol.

The combination of implicit solvation models with quantum mechanics was successfully used by many authors to predict solvation free energies. Sundararaman and Goddard published the CANDLE solvation model[57]. The benchmarked dataset consisted of 240 neutral molecules, 51 cations and 55 anions. They reported a mean unsigned error of 1.8 kcal/mol.

The Minnesota solvation database was frequently used for the parameterization of solvent models. Cramer and Truhlar who compiled this database used it to parameterize the SM*x*[47,48,49] models. For SM8 the mean unsigned errors of solvation free energies in aqueous solutions are 0.55 kcal/mol for neutral molecules and 3.20 kcal/mol for ions[11]. They also reported mean unsigned errors for IEF-PCM (neutral molecules: 4.87 kcal/mol, ions: 12.40 kcal/mol) C-PCM (neutral molecules: 1.57 kcal/mol, ions: 8.40 kcal/mol) and PB (neutral molecules: 2.28 kcal/mol, ions: 4.00 kcal/mol) and COSMO (neutral molcules: 2.76 kcal/mol, ions: 8.90 kcal/mol) for the same dataset in this paper.

Klamt *et al* proved in a follow up work[124] that these errors of the PCM and COSMO based methods may be decreased to a level close to SM8 for the subset of neutral molecules. In their work they calculated the solvation free energies of the neutral molecules with the level of theory used in the parametrization of PCM and COSMO. This reduced the mean unsigned error of COSMO-RS to 0.48 kcal/mol.

Dupont, Andreussi and Marzari focused on the prediction of free energies of solvation of ions with the self- consistent continuum solvation (SCCS) model. They motivated this study[56] with the words "applications of SM*x*[47,48,49] or related models[125,126,127] that lead to mean absolute errors below 1 kcal/mol (often around 0.5 kcal/mol) for neutral molecules give errors for charged species that are at least 4 kcal/mol." In their study they were able to improve the mean absolute error of the anions in the MNSol to 2.27 kcal/mol and the error of the cations' solvation free energies averaged at 5.54 kcal/mol.

As shown above, the coupling of quantum mechanics to implicit solvent models is on a promising trail towards chemical accuracy. However the development is focused on continuum solvation models and derivatives thereof. The reference interaction site model (RISM) is an alternative implicit solvation model which provides the user a different set of information than the continuum models. RISM gives access to the solvent distribution functions and is well suited for the prediction of relative free energies[5, 19,128, 24]. However absolute free energies

are usually inaccurate due to inherent approximations within RISM. Recently, quite some work has been done by different groups to improve the free energies of solvation calculated with RISM. A promising approach is the semi-empirical correction of the excess chemical potential based on the partial molar volume. The partial molar volume is scaled with an adjustable factor and eventually a constant offset. This was first done by Palmer, Frolov, Ratkova and Fedorov in 2010[129]. They showed for a test set of 120 organic molecules that the prediction error (μ^{RISM} - μ^{exp}) depends linearly on the partial molar volume. They proposed a correction model with a standard deviation of 0.76 kcal/mol on a training set and 0.99 kcal/mol on a test set. In a follow up paper[130] the authors showed that this correction also works on other datasets. They used a set of 21 druglike molecules and predicted their solvation free energies with the semi-empirical correction trained a set of 163 neutral molecules. These results show the universal nature of the correction, since the model was able to extrapolate to the on average lower solvation free energies of the druglike molecules.

Truchon, Pettitt and Labute further investigated the universality of the correction. They showed that the Ng[131] closure can be used instead of PSE-1. The also showed that the optimal coefficients are similar for LJ spheres and organic molecules. However they found that a reparametrization was necessary since they used a bigger data set (504 organic molecules) and a different water model.

Sergiievskyi *et al*[132,133,134]introduced a thermodynamic ensemble correction for RISM based on the partial molar volume. This parameter free correction is based on the addition of the transfer free energy from the grand canonical to the isobaric-isotherm ensemble. With this correction they were able to improve the quality of solvation free energies of 500 organic compounds calculated with RISM in comparison to experimental and simulation data. However there is a debate about correctness of the underlying thermodynamics.

Kinoshita et al studied the quality of partial molar volumes of amino acids calculated with the reference interaction site model[81]. They found that the partial molar volume is well described by 3D RISM/HNC and 1D RISM/HNC-BF0. In the case of 1D RISM the HNC-BF0 closure greatly increases the quality of the partial molar volume calculations.

7.2 Computational details

In this study a semi-empirical correction of free energies of solvation calculated with EC-RISM is developed. Towards this goal the findings of the earlier chapters will be used to identify a solid physical basis for this correction.

As shown in chapter 6 the excess chemical potential calculated using the reference interaction site model can be corrected by a bridge dependent functional. The free energy density of this functional is negative at small separations and rapidly vanishes beyond the σ_{LJ} for the Lennard Jones fluid. The bridge function inside the core has proven to have little impact on the HNC functional of the excess chemical potential. The in-core contributions to the excess chemical potential are located in the bridge dependent functional. Unfortunately the bridge functions of many relevant systems are still unknown. The bridge function can be approximated by various closures.

In this work the PSE-*k* closures are used and the bridge dependent functional is approximated using a semi-empirical correction. The free energy density is approximated by the direct correlation function, because this function is relative short ranged and has non-zero values inside the core. The direct correlation function is already calculated within RISM. This leads to a correction that can be rapidly evacuated.

As shown in equation (4.43) the partial molar volume can be calculated from an integration of the direct correlation function. Consequently this quantity is a well suited to approximate the bridge dependent functional. The approach chosen in the work is very similar to the approach by Palmer *et al.* and Truchon *et al.* which are described on page 66.

7.2.1 Model formulation of the partial molar volume correction

With the previously mentioned facts free energy correction models can be proposed. Plausible formulations are shown in table 8. Here models with up to 9 linear coefficients *c* will be used.

The minimal model (P2) has two adjustable parameters. One is used to fit the free energy contribution of the partial molar volume (c_V). In chapter 7.3.2 we will see that anions and cations need another adjustable parameter. Hence a second parameter c_q is added to fit charge contributions to the free energy of solvation. This parameter replaces the charge contributions to the partial molar volume[135]. Since the effect is the same magnitude and of opposite sign for anions and cations one parameter for all sorts of ions should be sufficient. The opposite sign will be introduced by the formal charge of the molecule.

To increase the flexibility of the model more parameters can be added. In the P3 model there is an adjustable constant (intercept) added. Hence this coefficient c_I is able to correct the free energy for errors not associated with the partial molar volume.

A variant of the P3 model is the P3NI model. Instead of an intercept, the excess chemical potential calculated by RISM is scaled with an adjustable parameter c_μ. In the case of the HNC functional this has a sound physical background. Since the HNC functional is evaluated on correlation functions based on incorrect bridge functions this coefficient can be used to compensate for this. If the PSE-*k* functional is used, the c_μ parameter can be interpreted as a long ranged correction for the incorrect bridge function, while the short ranged correction is applied by taking the partial molar volume into account using the c_V parameter.

Table 8: Semi-empirical correction models of EC-RISM free energies. Here μ^{RISM} is an excess chemical potential calculated with the reference interaction site model. In this work it can either be the HNC functional (4.32) or the PSE-*k* functional (4.41). The adjustable parameters are labelled with *c*. With the exception of c_I (which is an intercept) the subscript associates the parameter to a quantity. V_{M} is the partial molar volume, q is the formal charge of the molecule and E_{solv} is the electronic energy of the molecule in solution after the EC-RISM cycle. $\Delta\mu^{\mathrm{PSE}\text{-}k}$ is defined in equation (4.41). The free energy of solvation is given by $\Delta G_{solv} = \mu^{\mathrm{ex}} + E_{\mathrm{solv}} - E_{\mathrm{vac}}$. Due to the usage of the Ben Naim standard state[17] (see chapter 3.1) the ideal gas term in equation (3.1) cancels.

name	model
P2	$\mu^{\mathrm{ex}}_{\mathrm{P2}} = \mu^{\mathrm{RISM}} + c_V V_{\mathrm{M}} + c_q q_{\mathrm{Mol}}$
P3	$\mu^{\mathrm{ex}}_{\mathrm{P3}} = \mu^{\mathrm{RISM}} + c_V V_{\mathrm{M}} + c_q q_{\mathrm{Mol}} + c_I$
P3NI	$\mu^{\mathrm{ex}}_{\mathrm{P4NI}} = c_\mu \mu^{\mathrm{RISM}} + c_V V_{\mathrm{M}} + c_q q_{\mathrm{Mol}}$
P4	$\mu^{\mathrm{ex}}_{\mathrm{P4}} = c_\mu \mu^{\mathrm{RISM}} + c_V V_{\mathrm{M}} + c_q q_{\mathrm{Mol}} + c_I$
P4G	$\mu^{\mathrm{ex}}_{\mathrm{P4G}} = (1 + 10^{-8} c_G)(\mu^{\mathrm{RISM}} + E_{\mathrm{sol}}) + c_V V_{\mathrm{M}} + c_q q_{\mathrm{sol}} + c_I - E_{\mathrm{sol}}$
P4PSE	$\mu^{\mathrm{ex}}_{\mathrm{P4PSE}} = c_\mu \mu^{\mathrm{HNC}} + c_{\mathrm{d}\mu} \Delta\mu^{\mathrm{PSE}-k} + c_V V_{\mathrm{M}} + c_I$
P5	$\mu^{\mathrm{ex}}_{\mathrm{P5}} = c_\mu \mu^{\mathrm{HNC}} + c_{\mathrm{d}\mu} \Delta\mu^{\mathrm{PSE}-k} + c_V V_{\mathrm{M}} + c_q q_{\mathrm{Mol}} + c_I$
P9	$\mu^{\mathrm{ex}}_{\mathrm{P9}} = \begin{cases} c_{\mu,0} \mu^{\mathrm{RISM}} + c_{V,0} V_{\mathrm{M}} + c_{I,0} & \text{neutrals} \\ c_{\mu,-} \mu^{\mathrm{RISM}} + c_{V,-} V_{\mathrm{M}} + c_{I,-} & \text{anions} \\ c_{\mu,+} \mu^{\mathrm{RISM}} + c_{V,+} V_{\mathrm{M}} + c_{I,+} & \text{cations} \end{cases}$

The P4 model is a combination of both of the P3 models.

The P4G model takes the electronic energy of the solute into account. Since this value is usually orders of magnitudes higher than the excess chemical potential, small parameter changes have a huge effect on the result. To increase the numerical stability of the fitting process, the coefficient is not directly used to scale the sum of excess chemical potential and electronic

energy. It is used within the term $(1+10^{-8}\ c_G)$.This model may be able to compensate for errors in the electronic energy during the EC-RISM cycle.

The P4PSE model does not use the solvent charge as a parameter. Instead the PSE-*k* functional is used. Since the PSE-*k* closures are known for a good representation of long ranged interactions, this functional may be able to discriminate between anions, cations and neutral molecules. In this case one is likely to end up with a more predictive model because the PSE-*k* functional is certainly more unique of each solvent molecule than the solvent charge. The resulting model is assumed to be less dependent of the training dataset.

In the P5 model the combination of the charge term with the PSE-*k* functional leads to a very flexible model.

The P9 model is similar to the P4 model, but instead of a constant charge dependent term, it uses different parameter sets for anions, cations and neutral molecules.

Each EC-RISM free energy of solvation model will be parametrized using the levels of theory B3LYP[136,137], B3LYP-D3[138] and MP2[139]. The basis sets used are 6-31+G*, 6-311+G**[140] and aug-cc-pVDZ[141,142]. The 3D RISM part of the EC-RISM calculations will be performed with the closures PSE-1, PSE-2, PSE-3 and HNC-BF0. The susceptibility function which defines the solvent model was calculated from SPC/E water using 1D-RISM/HNC or MD-simulations (see chapter 7.2.3 for details). In the case of 1D-RISM/HNC the MSPC/E model (see chapter 3.3.1) was used instead. When the HNC-BF0 closure was used, the susceptibility function was calculated using 1D RISM/HNC-BF0. The combinations of these settings are shown in table 11. In the work (especially in the figures) the Pople basis sets 6-311G** and 6-31G* will be referred to as P-TZ and P-DZ while the Dunning basis set aug-cc-pVDZ will be called D-DZ.

In this work, two variants of the models P2, P3, P3NI, P4, P4G and P9 will be discussed. Each one uses another approximation towards the excess chemical potential calculated by RISM (μ^{RISM}). In the first variant it is assumed that the real bridge function is close to the PSE-*k* bridge function. In this case, the excess chemical potential is the sum of the HNC functional (μ^{HNC}, equation (4.32)) of the excess chemical potential and the bridge dependent term, which then is approximated by the scaled partial molar volume. In the second variant, the PSE solution is taken as given and the scaled partial molar volume is used to correct the theory itself. Then the PSE functional ($\mu^{\mathrm{PSE}\text{-}k}$, equation (4.41)) of the excess chemical potential is used and corrected by the scaled partial molar volume.

An important part of the calculation of the solvation free energy is the energy of the molecule in the gas phase. This will be called the reference energy. In this work the reference will be calculated with two approaches. In the first approach the geometries are optimized in gas phase and *all relevant* local minimum energy conformations are taken into account. This will be called vacuum reference geometries in this work. The usage of vacuum reference geometries is from the physics perspective the optimal way. However this can be time consuming. As an alternative the molecule geometry may be assumed the same optimal geometry in the gas and solvent phase. Then any structural reorganization contributions to the solvation free energies are ignored, but the calculation of a set of gas phase geometries is not needed. In the assumption the geometry of the solvated molecule itself is used as reference. Consequently this approach will be called "self reference" geometries.

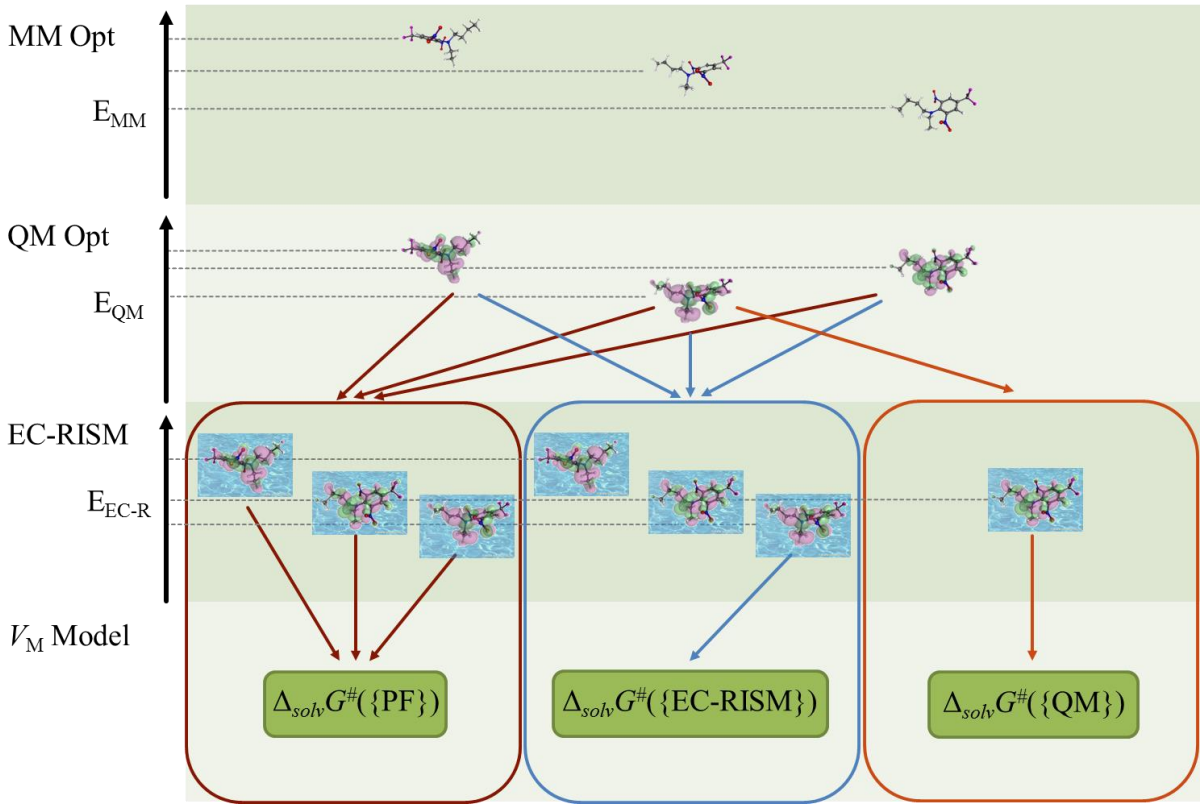


Figure 17: Ensemble averaging methods. **Red**: Partition functions. Here the complete set of geometries is used in EC-RISM and for the free energy calculation with the partial molar volume correction. **Blue**: EC-RISM minimum. The EC-RISM calculations are performed with all conformations. The partial molar volume correction is only applied to the lowest EC-RISM free energy. **Orange**: QM Minimum. The EC-RISM calculations are only performed with a single geometry. The one with the lowest QM energy is chosen.

For flexible molecules there are conformational contributions to the solvation free energy. To handle these contributions an ensemble of conformations has to be calculated and an averag-

ing has to be performed. However, by doing so the vibrational and rotational contributions to the free energy are still ignored. In the work this ensemble is built in three different ways. Firstly, the conformational contributions are neglected and the optimal geometry is assumed to be the same for EC-RISM and the previous optimization in PCM[43] solvent or vacuum. In this case the EC-RISM calculations are only done with the global minimum energy structure from the QM. This ensemble will be called "PCM". The second approximation is neglecting conformational contributions and assuming that the global minimum energy structure from the previous optimizations is different from EC-RISM. Therefore in the EC-RISM ensemble all local minima are used in the EC-RISM calculation and only the lowest energy from this step is considered. The third way is the usage of partition functions (PF) as described in equation (3.2) to consider the conformational contributions. This approach is similar to the EC-RISM ensemble, but instead of using the global minimum only, the contributions of all local minima are taken into account. An overview of the averaging methods is given in figure 17.

In the EC-RISM and PF ensembles there is a difference between the self and the vacuum reference. When vacuum geometries are used, the ensemble averaging is done first and the reference energy is subtracted. When the self reference model is used, the energy is subtracted before applying the ensemble averaging. In the case of vacuum reference geometries the conformational ensemble ("PF","EC-RISM","PCM") was generated in the same way the ensemble was generated in solution.

For the parametrization strategy EC-RISM and PF averaging have the disadvantage that they make the optimization process a non-linear problem. Hence in this work all optimizations were performed numerically as described in chapter 7.2.4. For example the minimization of the P2 model in the PF ensemble and with gas phase reference geometries the is given by

$$\begin{aligned} G_{solv} &= -RT \ln \sum_C \exp[-(E_C^{\mathrm{sol}} + \mu_C^{\mathrm{RISM}} + c_V V_{\mathrm{M,C}} + c_q q_{\mathrm{Mol}})/RT], \\ G_{vac} &= -RT \ln \sum_C \exp[-(E_C^{\mathrm{vac}})/RT], \\ &\min\{(G_{solv} - G_{vac})^2 - (\Delta_{solv} G_{\mathrm{MNSol}})^2 \,|\, c_V, c_q\}. \end{aligned} \tag{7.1}$$

The mathematical formulas of all models are shown in table 8. These models are trained with the experimental free energies of solvation of the Minnesota solvation database $\Delta_{slov}G_{\mathrm{MNSol}}$.

To increase the readability a shorthand notation will be used. The conditions under that a model was generated will be shown in an identifier like:

Model(Geom$_y^x$, QM/*Theory*/*Basis Set*, *B*[closure], χ[susceptibility], *F*[functional], {ensemble})

The labels in the identifier have the meaning of

- Model: the mathematical expessions from table 8,
- Geom: Defines the geometries used in the calculations. The solvation state geometry is labeled with *x*. It is described in table 9. The vacuum reference state is labeled *y*. When the geometry is the same in the gas (self reference) as it is in the liquid phase it is labeled *l*. When vacuum optimized geometries are used instead, the label is *g*. Possible combinations of *x* and *y* are shown in table 10.
- QM: Defines the QM settings during the EC-RISM calculations,
- *B*: Defines the closure/bridge function used in the EC-RISM calculations,
- χ: Labels the susceptibility function used in the EC-RISM calculations,
- *F*: Defines the free energy functional used in the model. Possible choices are HNC and PSE referring to equations (4.32) and (4.41),
- {}: Defines the method used for ensemble averaging.

As an example the identifier

$$\text{P3NI}(\text{Geom}_{\text{g}}^{\text{PCM}}, \text{QM/MP2/aug-cc-pVDZ}, B[\text{PSE-3}], \chi[\text{MD}], F[\text{HNC}], \{\text{PF}\})$$

means that the P3NI model was trained with the Geom^{PCM} dataset, gas phase optimized reference energies, at the MP2 level of theory using a aug-cc-pVDZ basis set (short: D-DZ) with the PSE-3 closure and the simulated solvent susceptibility functions. In the model the HNC functional was used and the calculations were done with multiple conformations of the same molecule and a partition function was used for the ensemble averaging.

The Minnesota solvation database contains molecules with multiple rotatable bonds. Hence a workflow is needed that accounts for this flexibility. This can be done by calculation of *all relevant* local minimum energy conformations. It is likely that there are multiple local minimum energy conformations. Hence the ensemble has to be averaged.

In this work a workflow is used that produces a small number of conformations for every molecule which are (potentially) local minima on the potential energy surface. For the calculation of solvation free energies we need two sets of conformations since this energy is associated with the process of transferring a molecule from the gas phase into a solvent. Consequently a set of gas phase optimized conformations and a set of conformations optimized in solution are needed for every molecule.

The solvent optimized conformations are treated with EC-RISM. They are used together with the gas phase energies calculated from the gas phase optimized conformers to train the free energy of solvation model (see figure 19).

The first step in the development of the semi-empirical model is a systematic screening of parameters within EC-RISM calculations to identify the optimal setup for free energy calculations. Then the performance of several free energy correction models which again depend on multiple parameters is tested.

Hence this work is based on a high dimensional nonlinear optimization process. The parameters to be optimized are:

- conformer generation
    - starting points of atom coordinates,
    - geometry optimization conditions in gas phase and solution,
    - selection of representative conformers,
- EC-RISM quantum mechanics
    - level of theory,
    - basis set,
- EC-RISM integral equation
    - closure,
    - solvent model,
- partial molar volume correction of the free energy
    - calculation of the partial molar volume,
    - empirical fit coefficients of the model,
    - conformer free energy contribution averaging.

The workflow and the parameters included will be discussed in detail in the following sections.

7.2.2 Conformer generation

The first step in the workflow towards the calculation of free energies is the generation of conformers as initial guesses. The goal is to get diverse sets of conformations without the need of extensive calculations. The ideal result of the workflow yields at least one conformation-guess close to all real relevant local energy minimum conformation. These guesses then can be optimized with a more elaborate theory.

In this work the conformations were generated by using three dimensional embedding based on distance geometry[143,144,145]. This means the conformations are built up by an optimization of random distance matrices. In this approach the possible maximum and minimum distances are constrained towards standard distances in molecules (see chapter 9.1).

The Minnesota solvation database provides M06-2X/MG3S optimized geometries for all molecules. These geometries were used to generate the topology of all molecules. The molecules in the Minnesota solvation database have up to 30 rotatable bonds. Therefore an exhaustive search of all conformations is not possible. As a compromise between computational effort and thorough scan of the conformational space 200 conformations were generated for each molecule using a distance geometry embedding[143] using the default settings of RDKit[3].

The next step is the optimization of the geometry of each conformer. This step has to be done twice: once in solution to prepare the EC-RISM calculation and once in vacuum to get the energy of the molecule in the gas phase.

In this work a scripted workflow by P. Kibies[146] was used for the classical mechanics and force field based minimization. These conformations were optimized using the GAFF[147,148] force field 1.5 of AMBER 12[41] and AM1-BCC charges. In the case of the molecules in solution the optimizations were done in ALPB[46] water. The energy minimizations were performed down to a maximum gradient norm of 10^{-4} kcal mol^{-1} $Å^{-1}$.

Since a quantum mechanics treatment of 200 conformations of over 500 molecules is too demanding, a filtering for representative conformers was needed at this point. Initial tests with various clustering algorithms implemented in g_cluster from the GROMACS 4.6.3 software suite[149] did not provide a consistent and stable setup. All tested settings yielded either one single cluster for most molecules or one cluster for each conformation of most molecules.

Therefore instead of a clustering based approach another algorithm was used to select possible minima. The force field minimized structures were sorted by their force field energy and the lowest energy conformation is considered the first local minimum. Now the RMSD to the next higher energy conformation is calculated. If this has an RMSD bigger than 0.1 Å this conformation is considered a new local minimum. Then this is repeated for all conformers with an energy not greater than the energy of the lowest energy conformer plus 5 kcal/mol. A conformer is considered as a local minimum when its RMSD to every other local minimum structure is bigger than 0.1 Å.

[3] http://www.rdkit.org/

The selected minima were then further optimized using quantum mechanical energy minimization. The optimizations were performed with Gaussian 09[150]. Every conformation was optimized under different conditions resulting in four datasets. They are shown in table 9.

Table 9: Datasets used for the parametrization of the partial molar volume correction. Each dataset consists of all molecules in the MNSol (for aqueous solutions) optimized under using different QM levels of theory. The $Geom^{MNSol}$ dataset was provided by the database[18]. These optimizations were done after the force field minimizations were performed in ALPB water. For the available reference geometries (force field minimizations in the gas phase) see table 9.

name of the dataset	optimization conditions
$Geom^{0}$	B3LYP/6-311+G**
$Geom^{PCM}$	B3LYP/6-311+G**/PCM
$Geom^{D3}$	B3LYP-D3/6-311+G**/PCM
$Geom^{DZ}$	B3LYP/6-31+G*/PCM
$Geom^{MNSol}$	M06-2X/MG3S

With all datasets a final filtering step was performed similar to the one after the force field minimizations. This time the energy cutoff was 3 kcal/mol. For the PCM optimizations the default settings for water were used. After this second filtering step there were still some molecules left with more than 100 conformations. This was considered as too many conformations of EC-RISM calculations. Hence for all molecules only the five lowest energy conformations were used in the subsequent EC-RISM calculations. With a cap at five clusters, 80% of the molecules of the MNSol have all identified conformers included into the EC-RISM calculations. Additions of further molecules will not significantly raise this percentage. This is shown in figure 18.

A fifth dataset was included as an external reference. The conformations provided by the Minnesota solvation database were used for this purpose. These geometries were previously used for the training of semi-empirical solvation models with the Minnesota solvation database[11] by other groups. Hence this dataset is well suited for the comparison of this model with published ones. It will be referred to as $Geom^{MNSol}$.

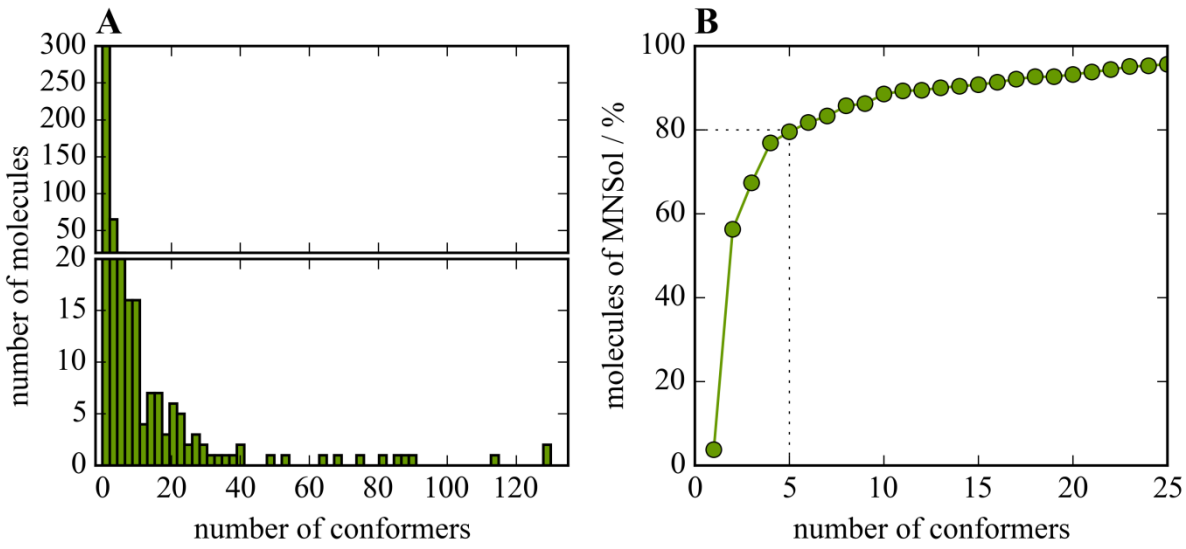


Figure 18: Conformer distribution in the MNSol. Panel **A**: The histogram shows the number of molecules for every amount of conformations. Panel **B**: The integral of the histogram shows the percentage of molecules that have equal or less conformers than the number on the x-axis.

The gas phase optimizations were not performed for all datasets. The "*g*" labeled datasets, which are physically more consistent (see the model identifier on page 71), are only available for $Geom^{0}$ and $Geom^{PCM}$. An overview of the datasets is shown in table 10. A schematic overview of the workflow is shown in figure 19.

Table 10: Combinations of geometry optimization (dataset) and reference geometries used in this work. The "*l*" datasets used the same geometry in solution and as a reference. These geometries were optimized in with a force field in ALBP water the further optimized as shown in table 9. When compared with figure 19 this corresponds to the blue and purple path. The "*g*" datasets used force field ALPB optimizations followed by a QM optimization as described in table 9 in solution and force field gas phase optimization followed by an optimization as a reference. The QM optimization was done with the same level of theory and basis set but without PCM. This is shown by the blue and red path in figure 19.

$Geom_l^0$	$Geom_g^0$
$Geom_l^{PCM}$	$Geom_g^{PCM}$
$Geom_l^{D3}$	
$Geom_l^{DZ}$	
$Geom_l^{MNSol}$	

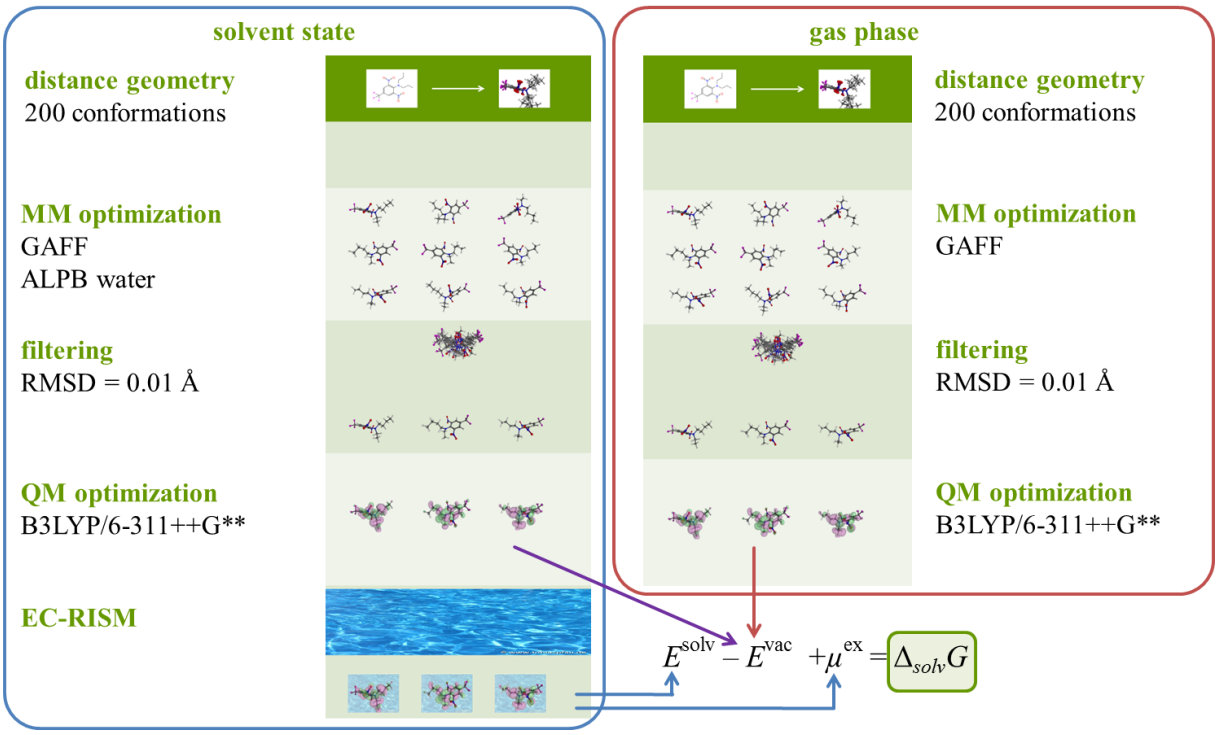


Figure 19: The workflow used to calculate solvation free energies. The left site shows the optimization of the molecules in solution and the right side shows the generations of gas phase references. The QM optimization step was performed at four different levels (see first four rows of table 9). After the QM optimization an additional filtering was performed as described on page 75 limiting the number of conformations used in EC-RISM to five. The reference energies of the $Geom_g$ datasets were calculated according to the red arrow. The calculations with self references ($Geom_l$) used the gas phase optimized structures. In this case the gas phase energy E^{vac} was calculated directly from the optimized geometries on the left side (purple arrow).

7.2.3 EC-RISM

With every geometry dataset, EC-RISM calculations were performed for the five lowest energy conformations of each molecule. The EC-RISM calculations were done with a set of fixed parameters and a set of parameters that was used to screen for optimal calculation conditions. The quantum mechanics part of the EC-RISM cycle was done with Gaussian03[151].

In all EC-RISM calculations the solvent density was set to 0.03333 $Å^{-3}$ and the temperature was adjusted to 298.15 K. The compressibility κ of water calculated be to 0.450183 Å/J (0.450183 10^{-9} Pa^{-1}). This value was calculated with the formula

$$\kappa = \frac{1}{\rho}\frac{\partial \rho}{\partial p}. \qquad (7.2)$$

The pressure p dependent water density ρ was taken from [152]. The ε_{LJ} and σ_{LJ} parameters of the solvent molecule's atoms were taken from the GAFF[147,148] force field 1.5. The solute solvent potential was calculated utilizing Lorentz-Berthelot mixing rules. Partial charges were

calculated from the quantum mechanics electrostatic potential using the CHelpG[153] scheme. The radius of Br atoms was set to 1.3 Å.

The RISM calculations were considered converged when the maximum residual of the direct correlation function fell below 10^{-5} for two consecutive iterations. A modified direct inversion of the iterative subspace (MDIIS)[111] was used to improve the convergence of the RISM/closure system of equations.

The initial guess of the partial charges were generated using the target level of theory (see below and table 11). The iterations were performed at HF level of theory. The EC-RISM iteration was stopped when the change of the sum of quantum mechanics solvent energy and the excess chemical potential fell below 0.01 kcal/mol. The iteration was performed on a cubic uniformly spaced grid with a point, point distance of 0.3 Å and a 12.5 Å solvent buffer in each direction. After the final iteration the grid was extended to 15 Å for a final evaluation of the excess chemical potential.

The parameters which were screened were the final level of theory of the EC-RISM calculation where MP2, B3LYP and B3LYP-D3 were used. The D3 corrections to the B3LYP calculations were applied with the program dftd3[138,154] using BJ-damping[155]. Three different basis sets were tested: aug-cc-pVDZ, 6-311+G** and 6-31+G*. The closures were set to PSE-1, PSE-2, PSE-3 or HNC-BF0. The solvent susceptibility function of water was calculated in three different ways: with 1D DRISM[70,71]/HNC, extracted from MD simulations and with 1D DRISM/HNC-BF0. An overview of all combinations is shown in table 11.

The first susceptibility function (HNC) was created using the DRISM/HNC method. The water model used in this case was a modified version of the SPC/E water model[36] with small σ_{LJ} parameters assigned to the hydrogen atoms $\sigma_{LJ}(H) = 1.0$ Å[37,38]. The dielectric constant was set to 78. A modified direct inversion of the iterative subspace (MDIIS)[111] was used to improve the convergence of the RSIM/closure system of equations. Convergence was assumed when the maximum residual of the direct correlation function fell below 10^{-8}. The calculations were performed on a logarithmically spaced grid[110] ranging from 0.0059 Å to 164.02 Å. This solvent susceptibility function was the same as used in [19].

The susceptibility function of the MD solvent model was calculated from a trajectory of SPC/E. The simulations and the calculations of the pair distribution functions were performed by D. Horinek. The trajectory built from 50000 snapshots of a 20 ns MD simulation in GROMACS[156] version 4.6.5. The simulations were performed with 17440 SPC/E water mol-

ecules at 298.15 K and 1 bar. A stochastic velocity rescaling thermostat was used in combination with a Parrinello-Rahman barostat. The simulations used a 6-12 Lennard-Jones potential with a 10 Å cutoff. A smooth particle-mesh Ewald summation[157] with a real space cutoff of 10 Å was used to calculate the electrostatic interactions. A time step of 1.5 fs was used and the water geometry was constrained using SETTLE[158]. The pair distribution function of water was calculated on an equally spaced grid of 1992 points with a grid point distance of 0.002 Å. It was calculated from a single precision trajectory.

The χ[MD] solvent susceptibility was calculated by S. Kast and P. Kibies using an iterative inversion of the OZ equation as described in chapter 5. The pair distribution function was interpolated on the logarithmic grid using a smoothing spline. The smoothing factor was set to the number of MD grid points and the statistical uncertainty was calculated from 195 samples of the 20 ns trajectory. The pair distribution function was set to zero when the grid point was closer to zero than 0.002 Å and is was set to the HNC solution for separations greater than the half of the box size of the MD simulation. The transition from the MD simulation based closure to the HNC was done within 1 Å using a cubic polynomial. The convergence tolerance was increased to 10^{-4}.

The susceptibility of the HNC/BF0 water model was calculated in a two-step process. With the exception of the following settings the calculations were done in the same way das for the HNC solvent susceptibility function. In the first step all partial charges were set to zero. The Verlet closure was used with a convergence criterion of 10^{-5}. In the second step, the bridge function of the first step was used with a reference HNC approach. The charges of SPC/E water were reassigned to the atoms. The dielectric constant was set to 78.4. The bridge function from step one was added to the solute solvent interaction potential and a RISM/HNC calculation was performed with the same settings as used in the first step.

From every conformation of each molecule used in EC-RISM a 3D-RISM calculation was performed with all partial charges set to zero. These non-polar calculations were done with the same settings for 3D-RISM as used in the EC RISM calculations. The partial molar volume of the uncharged (non-polar) molecules was calculated as well.

Table 11: Number of successfully performed EC-RISM calculations using different EC-RISM settings and geometries for neutral molecules, cations, anions and the total number. This is an overview of all combinations of screened EC-RISM parameters used in this work for every geometry dataset (see table 9). When B3LYP is the level of theory, the D3 dispersion energy was calculated as well.

Geom	level of theory	basis set	closure	susceptibility function	neutrals	cations	anions	all
0	B3LYP	6-311+G**	PSE-1	HNC	339	45	73	457
				Sim	340	45	73	458
			PSE-2	HNC	339	45	73	457
				Sim	339	45	73	457
			PSE-3	HNC	339	43	73	455
				Sim	339	45	72	456
	MP2	6-311+G**	PSE-1	HNC	355	49	82	486
				Sim	363	49	82	494
			PSE-2	HNC	348	49	82	479
				Sim	345	45	73	463
			PSE-3	HNC	353	47	82	482
				Sim	350	47	74	471
		aug-cc-pVDZ	PSE-1	HNC	305	45	72	422
				Sim	323	44	78	445
			PSE-2	HNC	302	44	71	417
				Sim	319	35	63	417
			PSE-3	HNC	301	44	71	416
				Sim	321	29	74	424
D3	B3LYP	6-311+G**	PSE-1	HNC	368	47	78	493
			PSE-3	HNC	253	21	39	313
	MP2	aug-cc-pVDZ	PSE-1	HNC	255	47	77	379
			PSE-2	HNC	253	45	69	367
			PSE-3	HNC	313	43	69	425
DZ	B3LYP	6-31+G*	PSE-2	HNC	375	47	79	501
	MP2	aug-cc-pVDZ	PSE-3	HNC	327	47	76	450
MNSol	MP2	aug-cc-pVDZ	PSE-3	HNC	324	48	77	449

Table 11 continued

Geom	level of theory	basis set	closure	Susceptibility function	neutrals	cations	anions	all
PCM	B3LYP	6-311+G**	PSE-1	HNC	345	49	76	470
				Sim	374	50	80	504
			PSE-2	HNC	366	46	67	479
				Sim	369	47	76	492
			PSE-3	HNC	357	49	80	486
				Sim	356	50	79	485
	MP2	6-311+G**	PSE-1	HNC	352	50	80	482
				Sim	354	50	80	484
			PSE-2	HNC	355	50	80	485
				Sim	358	49	79	486
			PSE-3	HNC	355	49	80	484
				Sim	351	48	77	476
		aug-cc-pVDZ	HNC	HNC-BF0	211	5	5	221
			PSE-1	HNC	334	50	79	463
				Sim	325	49	78	452
			PSE-2	HNC	284	50	76	410
				Sim	329	50	79	458
			PSE-3	HNC	325	50	77	452
				Sim	326	50	77	453

7.2.4 Applying the partial molar volume correction models

As shown in table 8, several variants of a semi-empirical correction of free energies of solvation were tested in this work.

For every EC-RISM setup in table 11 (including the DFT-D3 corrections) all models in table 8 were parameterized. The model generation was done in several variants:

- μ^{RISM}
    - HNC functional and closure is one of PSE-*k* or VM dependent of the setup,
    - PSE functional only for setups with PSE-*k* closure,
- conformational ensemble averaging
    - global minimum of the EC-RISM calculations,

  - global minimum after the PCM optimization step,
  - partition function,
- gas phase reference
  - own geometry *l* (neglecting conformational changes at the transition from the gas phase to the solvent),
  - gas phase optimized geometries *g* (only $\mathrm{Geom}_g^{\mathrm{PCM}}$ and Geom_g^{0}).

The choice of the reference state determined the parametrization strategy. When the gas phase optimized geometries were used, the following steps were made:

- guess initial parameters ($c_\mu = 1$, $c_v = 1$, $c_q = 0$, $c_I = 0$),
- calculate gas phase energy of the molecule
  - $\mathrm{G}_{\mathrm{conf}}^{\mathrm{vac}} = \mathrm{E}_{\mathrm{conf}}^{\mathrm{vac}}$
  - ensemble average ($\mathrm{G}_{\mathrm{mol}}^{\mathrm{vac}} = \mathrm{PF}[\{\mathrm{G}_{\mathrm{conf}}^{\mathrm{vac}}\}]$) or global minimum ($\mathrm{G}_{\mathrm{mol}}^{\mathrm{vac}} = \min[\{\mathrm{G}_{\mathrm{conf}}^{\mathrm{vac}}\}]$) of all conformations. Since the Energy of the reference state is not dependent on the partial molar volume correction this can be pre-calculated and hence does not affect the linearity of the optimization.
- calculate free energy in solution
  - $\mathrm{G}_{\mathrm{conf}}^{\mathrm{solv}} = \mathrm{E}_{\mathrm{conf}}^{\mathrm{solv}} + \mu_{\mathrm{conf}}^{\mathrm{ex}}$
  - calculate $\mu_{\mathrm{conf}}^{\mathrm{ex}}$ using any model in table 8
  - ensemble average ($\mathrm{G}_{\mathrm{mol}}^{\mathrm{solv}} = \mathrm{PF}[\{\mathrm{G}_{\mathrm{conf}}^{\mathrm{solf}}\}]$) or global minimum ($= \min[\{\mathrm{G}_{\mathrm{conf}}^{\mathrm{solf}}\}]$) of all conformations. Here the non-linearity of the optimization is introduced.
- calculate $\Delta G^{\mathrm{solv}}, = \mathrm{G}_{\mathrm{mol}}^{\mathrm{solf}} - \mathrm{G}_{\mathrm{mol}}^{\mathrm{vac}}$
- calculate total square of errors:

$$\Delta\Delta G^2 = \sum_{\mathrm{mol}} \Delta G_{\mathrm{calc}}^{\mathrm{solv}}[\mathrm{mol}] - \Delta G_{\mathrm{exp}}^{\mathrm{solv}}[\mathrm{mol}],$$

- minimize $\Delta\Delta G^2$ by variation of model coefficients.

When the conformation's own geometry was used, the following steps were made:

- guess initial parameters ($c_\mu = 1$, $c_v = 1$, $c_q = 0$, $c_I = 0$),
- calculate gas phase energy of the conformations of each molecule
  - $\mathrm{G}_{\mathrm{conf}}^{\mathrm{vac}} = \mathrm{E}_{\mathrm{conf}}^{\mathrm{vac}}$,
- calculate free energy in solution of each conformer
  - $\mathrm{G}_{\mathrm{conf}}^{\mathrm{solv}} = \mathrm{E}_{\mathrm{conf}}^{\mathrm{solv}} + \mu_{\mathrm{conf}}^{\mathrm{ex}}$,
  - calculate $\mu_{\mathrm{conf}}^{\mathrm{ex}}$ using any model in table 8,
- calculate ΔG^{solv},

- Calculate ensemble average or global minimum of all conformations
  - $\Delta G_{mol}^{solv} = PF[\{\Delta G_{conf}^{vac}\}]$ or $\Delta G_{mol}^{vac} = \min[\{\Delta G_{conf}^{vac}\}]$,
- calculate total square of errors
  - $\Delta\Delta G^2 = \sum_{mol} \Delta G_{calc}^{solv}[mol] - \Delta G_{exp}^{solv}[mol]$,
- minimize $\Delta\Delta G^2$ by variation of model coefficients.

The minimization was done using a BFGS algorithm. The molecules from table 12 were excluded from the optimization process.

Table 12: Some molecules of the Minnesota solvation database have diastereomers or E and Z conformers. In these cases the free energy of solvation depends on the relative conformation. To avoid errors these molecules were excluded from the parametrization.

database ID	reason
0020cis	multiple stereo centers and diastereomers
0028Epe	E / Z isomeres
0171Zdi	E / Z isomeres
0172Edi	E / Z isomeres
0234Emb	E / Z isomeres
0235Zmb	E / Z isomeres
0440pho	multiple stereo centers and diastereomers
c091	parametrization and convergence issues
i091	parametrization and convergence issues
test1014	multiple stereo centers and diastereomers
test1018	multiple stereo centers and diastereomers
test1019	E / Z isomers
test1029	multiple stereo centers and diastereomers
test1030	multiple stereo centers and diastereomers
test1033	multiple stereo centers and diastereomers
test1035	multiple stereo centers and diastereomers

7.3 Results

The partial molar volume correction was successfully parameterized with the Minnesota solvation database. This chapter summarizes and discusses the generation of the compounds, the parametrization and resulting models. At first the conformer generation and EC-RISM calculations are discussed. The information from the EC-RISM calculations is used to identify how

the partial molar volume correction can be applied and the effect of the solvent induced polarization is reviewed. In the second part the results of the partial molar volume correction are discussed. The overall performance measured by the RMSE and differences between neutral molecules and ions are identified. Then the settings of the QM calculations, the 3D RISM calculations and the correction itself are discussed. The correction is compared to other implicit solvation models and the robustness of the parametrization is checked. Finally the RMSE of substance classes is discussed and some general suggestions are given how the models can be used to minimize the expected error of EC-RISM calculations with the partial molar volume correction. In total 20372 QM minimizations (including all conformers) and 44083 EC-RISM calculations were performed. This resulted in 4228 different models and 3543960 free energies.

7.3.1 Conformer generation and geometry optimizations

The first step was the generation and force field optimization of the 533 molecules in the MNSol. For six molecules (test1030, i050, i051, i052, i053 and i088) the parametrization with the GAFF force field could not be applied automatically so these molecules were not included in any of the following steps. The first geometrical selection step yielded 265 molecules with only one conformation, 447 molecules with less than ten conformations. 40 molecules were left with at least 25 conformations. Nine iodine containing molecules were not used in the QM optimizations since iodine cannot be described with the basis set used in this work.

In the subsequent QM optimizations for 514 molecules one or more conformations were successfully optimized with B3LYP/6-311+G**. In case of B3LYP/6-311+G**/PCM it were 513 molecules, in case of B3LYP-D3/6-311+G**/PCM 508 molecules and in case of B3LYP/6-31G* 509 molecules.

The EC-RISM calculations were performed with the five lowest energy local minimum geometries of every molecule. The numbers of converged calculations for each calculation type are shown in table 11.

7.3.2 EC-RISM

This chapter gives an answer to the question how the difference from the solvation free energies calculated with EC-RISM to experimental values can be used to formulate a correction model.

So far the EC-RISM has proven to be useful for the prediction of changes in the free energy of solvation[5,19,128,24]. However, the prediction of absolute free energies of solvation with RISM has significant systematic errors. In figure 20 the calculated free energy of solvation is plotted against the experimental values from the of the MNSol database. The correlation of theory and experiment is low.

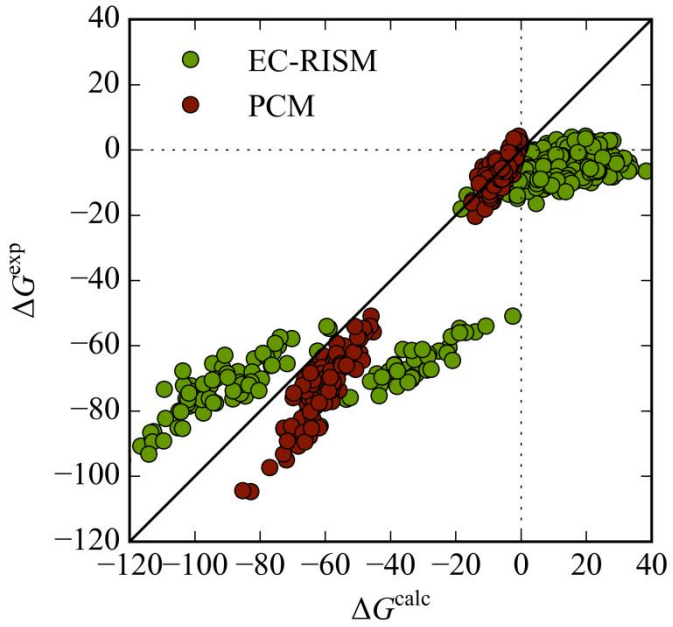


Figure 20: Calculated free energy of solvation plotted against the experimental values from the of the MNSol database. The EC-RISM calculations were performed with the Geom^0 dataset at the B3LYP/6-311+G**/EC-RISM-PSE-3 level with the HNC solvent model. The PCM was done at the same level. At this point the reader is reminded that PCM is able to perform much better on this dataset if the computational setting is chosen accordingly to the parametrization of PCM[124], which here is not the case.

An important question is how to compensate the error introduced by the EC-RISM calculation by employing the partial molar volume correction. A linear correlation between the difference of the calculated free energies and the experimental with the partial molar volume calculated by RISM is necessary for this. The correlation of the partial molar volume and the energy difference is shown in figure 21.

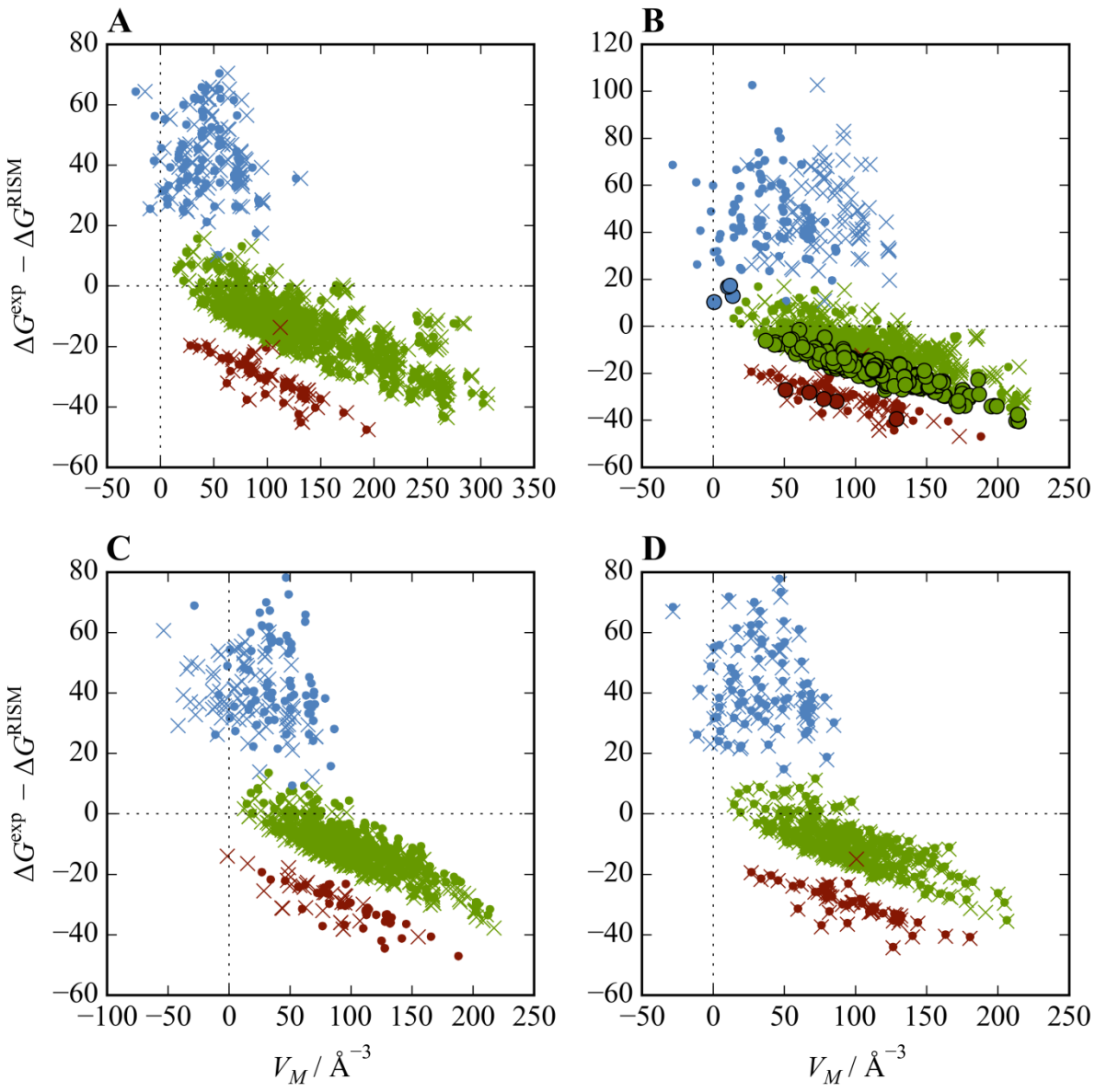


Figure 21: Plot of the partial molar volume against the error of EC-RISM calculations. Neutral molecules are labeled green, anions blue and cations red. All calculations use the PSE-*k* functional for the evaluation of the excess chemical potential. Figure **A** shows Geom^{DZ} dataset with B3LYP/6-31+G*/PSE-1(EC-RISM) with the HNC solvent model. The dots denote V_M of the final iteration, crosses denote V_M of the first iteration (vacuum). Figure **B** shows Geom^{PCM} dataset with MP2/aug-cc-pVDZ/PSE-3(EC-RISM) with the HNC solvent model. The dots correlate V_M of the final iteration, crosses denote V_M of the uncharged molecule, the circles display that the EC-RISM cycle was performed with the BF-0 closure. Figure **C** shows Geom^{0} dataset with MP2/aug-cc-pVDZ/PSE-3(EC-RISM). The dots are with the HNC solvent model and the crosses are calculated with the MD solvent susceptibility. Figure **D** shows the Geom^{MNSol} dataset with MP2/aug-cc-pVDZ/PSE-3(EC-RISM) with the HNC solvent model. The cross marked calculations use the HNC functional for the evaluation of the excess chemical potential. The dots mark usage of the PSE-k functional.

There is a multi-linear dependence of the error from the partial molar volume. Neutral molecules and cations have a much smaller variance within the data (in the y-direction) than the anions. Therefore a partial molar volume correction is expected to work better for cations and neutral molecules. The influence of the theoretical setup is rather small. Neither the free energy functional nor the susceptibility change the result significantly (panel C and D of figure 21). In case of PCM geometries (panel A and B of figure 21) the partial molar volumes are slightly bigger. The MD solvent susceptibility function decreases the partial molar volume slightly. The partial molar volume of charged molecules needs an additional term that depends only on the total charge of the molecule and not on any other properties[135]. Consequently the contributing of a monovalent cation is equal in magnitude and opposite in sign to the contribution of a monovalent anion. This is results in the multi-linear behavior in figure 20 and figure 21. Hence introducing a charge dependent term to the models has a solid physical background.

The calculations with the BF-0 closure (panel B of figure 21) had serious convergence issues, especially in case of the ions. Anyways this closure drastically decreases the y axis variance of the neutral molecules. At least for the neutral molecules, this closure may be a promising candidate. Since the accurate treatment of ions in the free energy correction model is a major goal of this work this promising path will only be followed shortly in chapter 7.3.11.

Using the non-polar partial molar volumes decreases the x-axis diversity of the data as seen in panel C of figure 21. Since these calculations require an additional RISM step and are unlikely to increase the quality of the resulting models this approach is discarded.

Figure 21 also shows that the partial molar volume is not able to discriminate between anions, cations and neutral molecules. The values of the anions tend to be a bit smaller. When looking only at the x axes in figure 21, one cannot see a clear cluster for the anions. Hence the charge of the molecule has to be included into the parameterization of the free energy correction.

So far[129,130,131] the partial molar volume correction was used in combination with 3D RISM only. This means that solvent induced polarization of the molecules is ignored. However the polarization energy ΔE^{Pol} ranges from -0.61 kcal/mol to 25.04 kcal/mol for neutral molecules and from -5.25 kcal/mol 67.17 kcal/mol for ions. For the partial molar volume correction to work, the partial molar volume has to over-compensate the lack of polarizability. This is possible, if the polarizability has a linear correlation with the partial molar volume. As seen in figure 22 this is partially the case. However the trend is different for anions. So the anions will

require a different parametrization. The correlation between partial molar volume and polarization energy is not optimal. Therefore calculation of the polarization energy with EC-RISM can improve a partial molar correction.

The polarization of a molecule is penalized due to the induced charge separation. As expected the penalty is highest for cations which tend to have high charge densities. The polarization energy during the transfer of a molecule into water should positive. However, in the MD solvent (panel **A** of figure 22) this value is negative of some cations and a meaningful amount of anions. This means that polarity of the molecule is decreased in the solvent, what appears to be an artificial event possibly introduced by the combination of DRISM with the simulated susceptibility functions.

7.3.3 Partial molar volume correction

The optimal model coefficients were calculated for 4228 combinations of computational parameters. Unfortunately in the 84 setups with the HNC/BF0 closure the iterations are likely to fail. Hence the amount of converged ions is very small in these cases. The HNC/BF0 calculations were therefore excluded in the following part and are discussed in chapter 7.3.11.

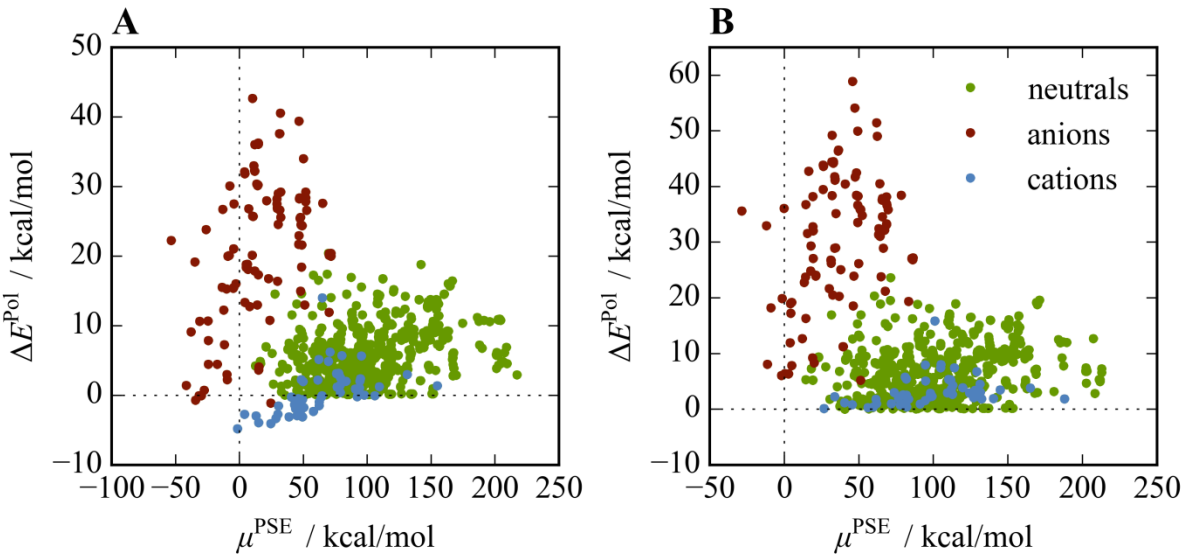


Figure 22: Polarization energy of the molecules in the MNSol plotted against the partial molar volume. The calculations were done with the Geom^{PCM} dataset using MP2/aug-cc-pVD/EC-RISM-PSE-3 with the simulated susceptibility function (**A**) and the HNC susceptibility function (**B**).

7.3.3.1 Overview

The RMSE of the optimal models for the complete Minnesota Solvation database range from 1.9 kcal/mol to more than 8 kcal/mol. The minimum and maximum RMSE for each model is

shown in table 13. The group containing the models P2, P3 and P4G as well as the group containing P3NI, P4, P5 and P9 models show similar performances within a group. The RMSE in the second group are about 0.2 kcal/mol lower. Cations are described better than anions. This is expected from the findings in chapter 7.3.2. Since the amount of anions and cations is smaller than the number of neutral molecules (see table 11), the overall mean error is close to the mean error of the neutral molecules.

Table 13: Overview of the highest and lowest RMSE (kcal/mol) of solvation free energies from the MNSol calculated for every model. For a comparison with continuum solvation models see chapter 7.3.5. More information about the lowest RMSE of each model is given in the appendix (The result tables. See chapter 9.3 and the following chapters). Here the setup corresponding to the RMSE is given as well. In most cases the MP2 level of theory yields the lowest RMSE.

	all molecules		anions		cations		neutrals	
model	min	max	min	max	min	max	min	max
P2	2.18	6.45	3.68	8.68	2.99	11.53	1.21	6.37
P3	2.18	4.82	3.67	6.02	2.71	7.39	**1.16**	4.06
P3NI	**2.04**	5.91	**3.45**	5.07	**2.53**	3.87	**1.18**	6.27
P4	**2.00**	3.94	**3.36**	4.94	**2.54**	3.59	**1.12**	3.96
P4G	2.18	4.82	3.67	6.02	2.68	7.38	1.16	4.06
P4PSE	3.00	8.03	5.19	12.16	3.61	20.04	2.02	4.89
P5	1.97	3.94	3.40	4.60	2.53	3.66	1.14	3.96
P9	**1.89**	3.85	**3.07**	4.56	**2.21**	3.40	**1.10**	3.95

By looking at the minimum RMSE for each model one can postulate some initial hypotheses about the effect of certain parameters. A comparison of the P2 and P3 model shows that the intercept has very little benefit. The minimal RMSE of neutral molecules decreases by 0.05 kcal/mol. In both cases, the computational setups are similar. Both use the $Geom^{PCM}$ dataset, the MP2/6-311+G** level of theory for the quantum mechanics, the PSE-1 closure, vacuum reference geometries and HNC functional for the energy evaluation. The added intercept benefits the simulated susceptibility function and the P2 model works better with the 1D RISM/HNC susceptibility. Table 14 shows that using this setup the intercept has almost no influence on the HNC RMSEs.

The P4 model has a significantly lower RMSE (RMSE(P4) = 2.00 kcal/mol) than the P3 model (RMSE(P2) = 2.18 kcal/mol). The added parameter to the excess chemical potential c_μ

seems to have a much bigger impact than the addition of an intercept. When comparing the P3NI model with the P4 importance of the c_μ is further supported. For the P4 and P3NI models the ideal settings are the same. In contrast to those models without this parameter, the PSE-3 closure and the PSE-3 functional for the energy evaluation are optimal. The simulated susceptibility performs better than the 1D RISM/HNC solvent model.

Table 14: RMSE of the P2 and P3 model calculated using the Geom^{PCM} dataset, the MP2/6-311+G** level of theory for the quantum mechanics, the PSE-1 closure, vacuum reference geometries and HNC functional for the energy evaluation.

susceptibility function	ensemble	P2	P3
HNC	EC-RISM	2.24	2.24
	PCM	2.24	2.24
	PF	2.23	2.23
Sim	EC-RISM	2.28	2.20
	PCM	2.28	2.20
	PF	2.30	2.22

The added five parameters to the P9 model when compared with the P4 model further increase the accuracy of the model. However the RMSE is only 0.11 kcal/mol lower than the RMSE of the P4 model. The P9 model does not improve the quality of the description of the neutral molecules but it has a great impact on the ions.

The P4G model has almost the same characteristics as the P3 model. This shows that the added parameter has no effect. The result is physically and numerically plausible. The P4G model scales the electronic energy of the molecule which is several orders of magnitudes higher than the excess chemical potential. Therefore the correction has either a huge impact or the almost no impact.

The P5 and the P4PSE model each include a separate parameter for the PSE-*k* functional $c_{d\mu}$. In the P4PSE model this parameter is supposed to compensate the lack of a charge parameter. This model fails in the description of ions. Since the number of cations in the MNSol database is much smaller than the amount of anions, the optimization leans towards a better description of the anions. Hence the cations have a much bigger RMSE what is therefore probably an artifact of the optimization and the unbalanced training database. Within the P5 model the $c_{d\mu}$ is used to add some more flexibility to the calculation. The improvement of the minimal RMSE is quite small, but the ideal setup changes. In the case of the Geom^{PCM} dataset, the MP2/6-

311+G** level of theory for the quantum mechanics, the PSE-1 closure, MD susceptibility, vacuum reference geometries and PSE-*k* functional for the energy evaluation (the optimal settings of the P4 model) cannot be used for the P5 model, since the $c_{d\mu}$ parameter needs evaluation of the HNC functional. So a direct comparison of the best models is not possible. The P5 model works well with the PSE-2 closure. Using the previously mentioned setup the decrease of the RMSE is 0.08 kcal/mol from P4 to P5 with the PSE-2 closure (RMSE(P5) = 2.01). For the PSE-3 closure the difference is only about 0.02 kcal/mol (RMSE(P5) = 2.02).

Figure 23 show the RMSE of the P2, P3, P4 and P9 of the full database (anions, cations and neutral molecules) calculated using various computational setups. These are the RMSE of the results of the optimization of the models.

At first one can see that an increase in the amount of parameters in the model will lower the RMSE of the model in all cases. The P4 and P9 model show consistent RMSE for almost all conditions, while the RMSE of P2 and P3 shows high variance over different levels. These out layers usually include an evaluation of the PES-*k* functional (labelled μ^{ex} in the figures). Also these models do not work well when the MD simulation solvent susceptibility (labelled Sim in the figures) function is used.

Also from figure 23 can be seen that the MP2 level of theory is usually better than B3LYP. It has lower RMSE differences between the models and the RMSE is less dependent of the other settings. This is a promising result. Since the quantum mechanics calculations are independent of the evaluation of the reference interaction site model, the partial molar volume correction is not expected to have an influence on the electronic energies. This means higher quality electronic energies calculations are supposed to improve the overall correction. If this is not the case the partial molar volume correction will over-compensate those errors. When B3LYP calculations are required, usage of the simulated solvent susceptibility functions with the P4 model is suggested for all geometries.

Figure 24 shows the P3, P3NI, P4G, P4 and P5 models. The P3 and the P4G model have similar RMSE for all setups. Between the P4 and P3NI models a similar relation exists. The P5 model has slightly lower RMSE than the P4 model. The difference between these models is relatively big when the simulated solvent susceptibility function is used in combination with MP2 and the PSE-1.

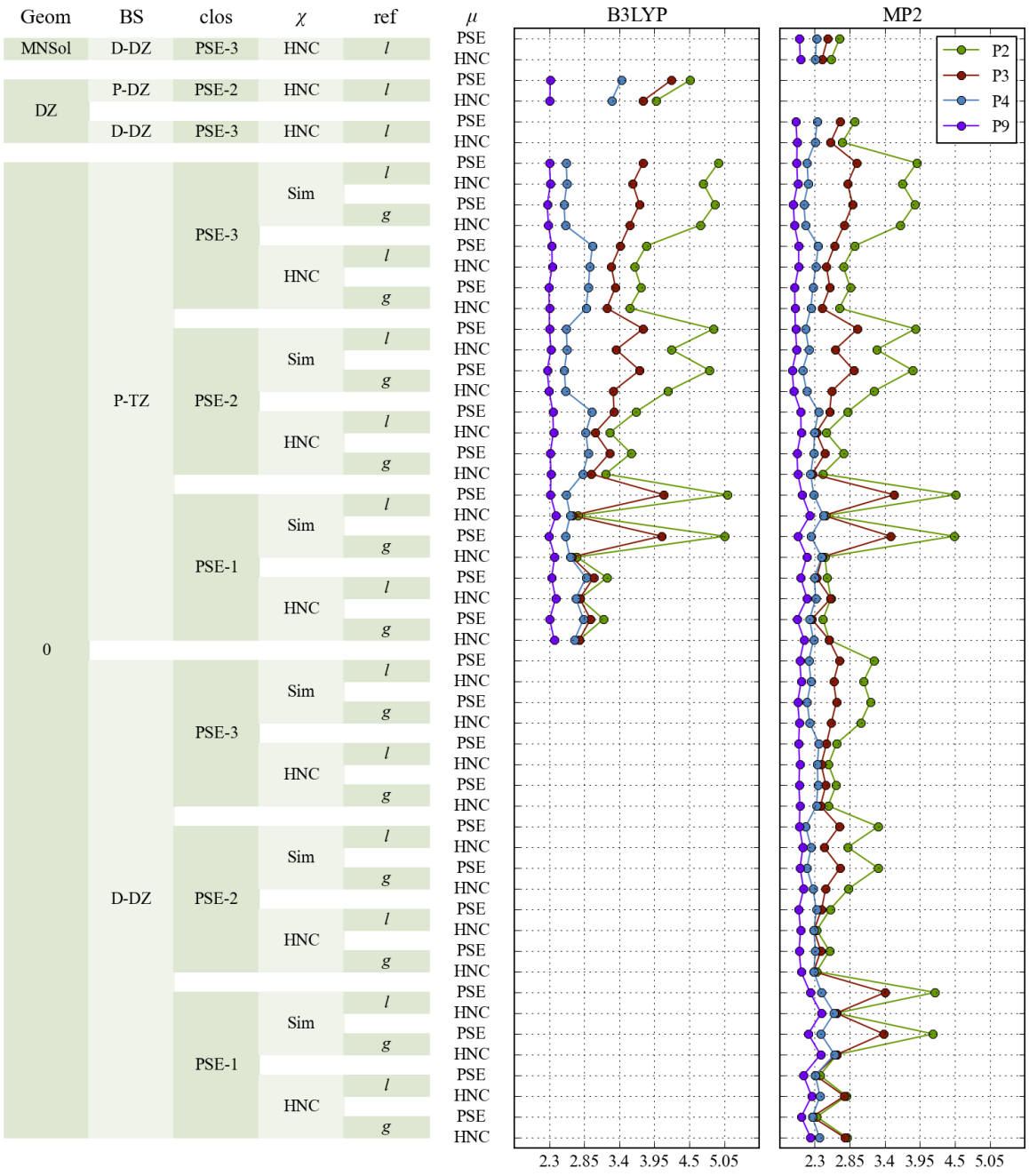


Figure 23: RMSE in kcal/mol of the free energy of solvation of the MNSol calculated with variants of the partial molar volume correction. This figure includes all generated variants for the P2, P3, P4 and P9 model with the EC-RISM ensemble, excluding EC-RISM calculations with the Verlet closure and calculations with $Geom^{D3}$/B3LYP-D3. Most of the later figures focus on subsets of this plot. The identifier corresponding to the shown models is Model($Geom^{Geom}_{ref}$, QM/LoT/BS, B[clos], χ[χ], F[μ], {EC-RISM}).

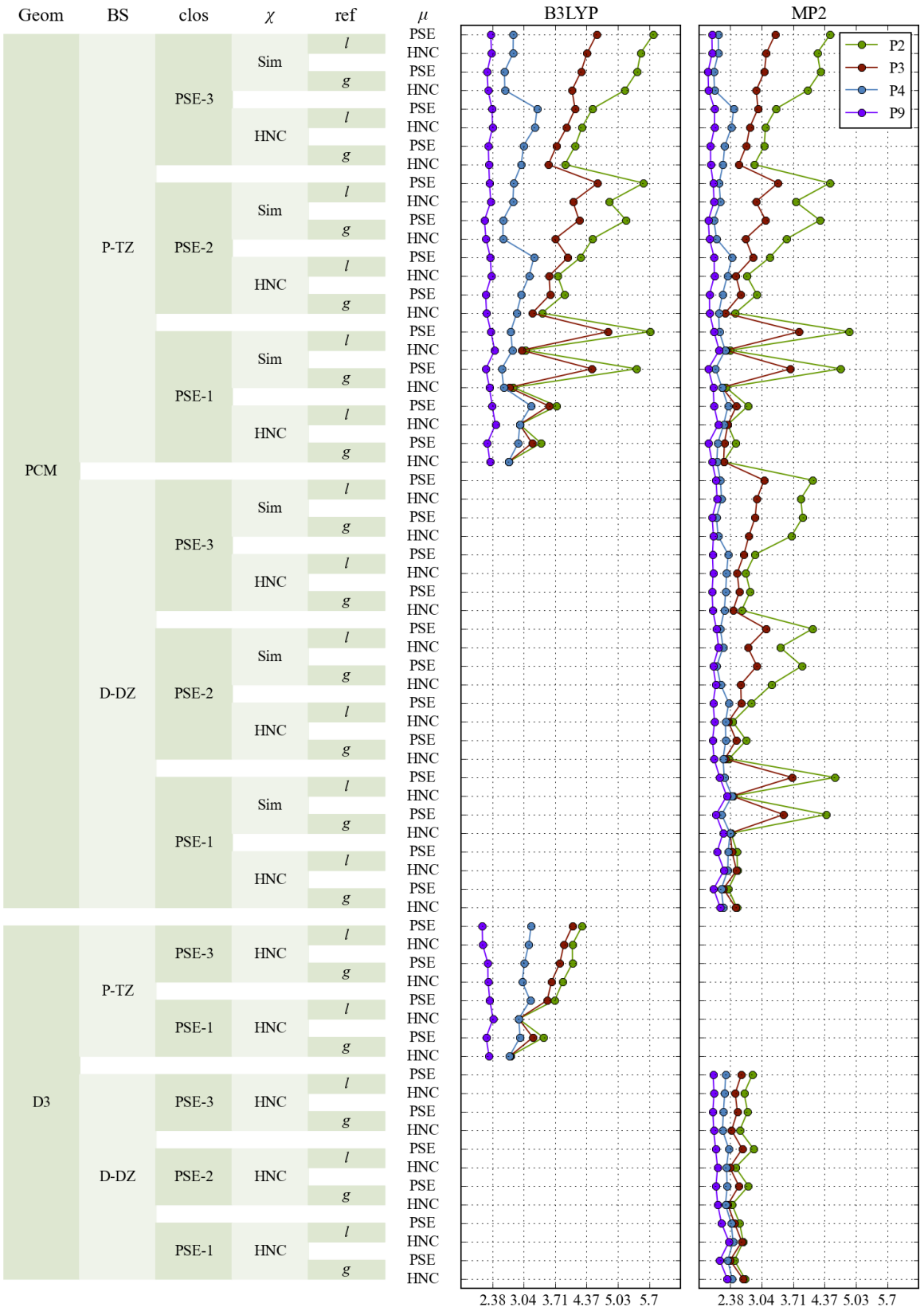


Figure 23 continued

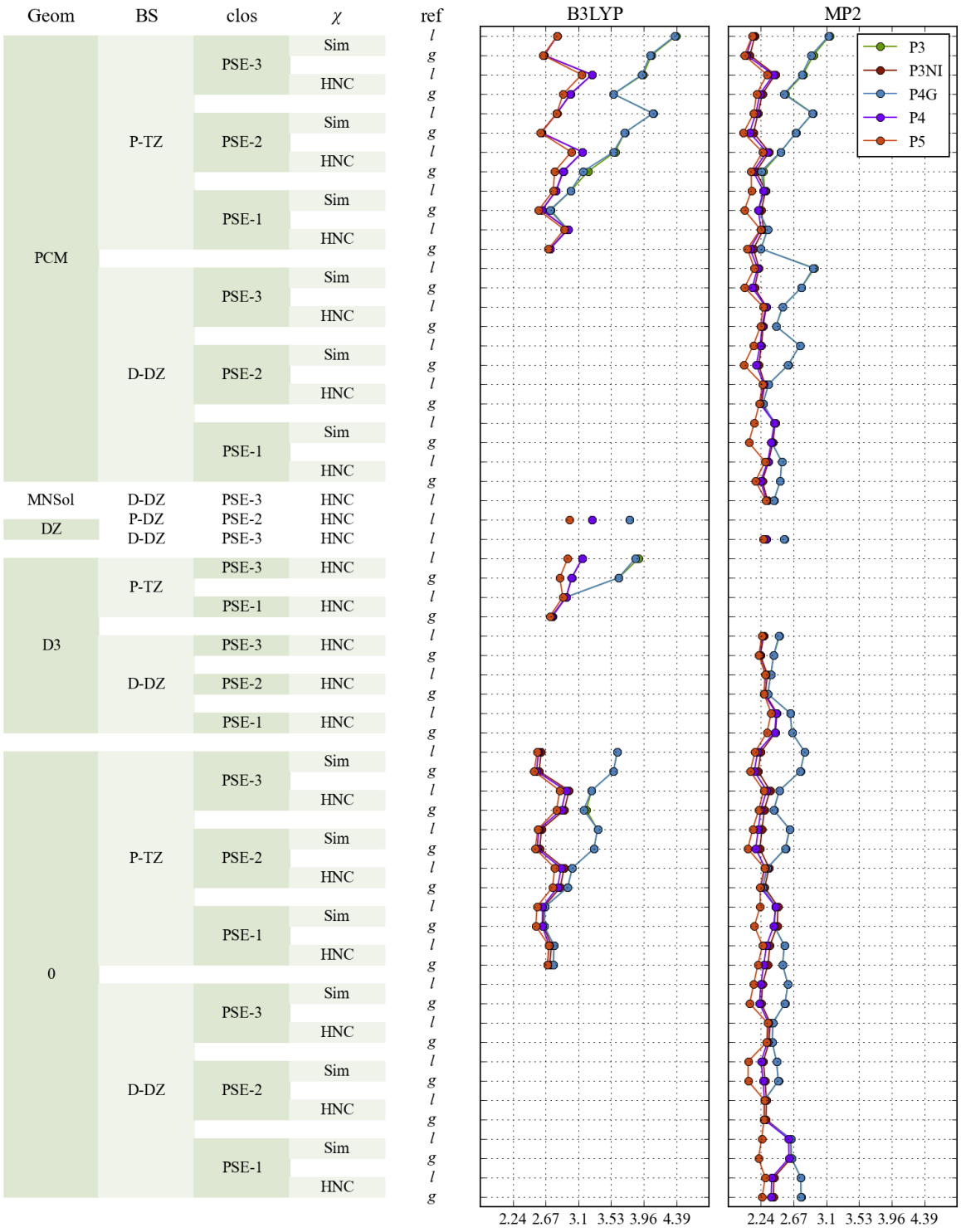


Figure 24: RMSE in kcal/mol of the free energy of solvation of the MNSol calculated for the P3, P3NI, P4Gm P4 and P5 models with the identifier Model($\text{Geom}_{\text{ref}}^{\text{Geom}}$, QM/LoT/BS, *B*[clos], χ[χ], *F*[HNC], {EC-RISM}). All models use the HNC functional since it is required for the P5 model.

7.3.3.2 Ionic and neutral molecules

The RMSE of the overall dataset can be separated into a subset for the anions, cations and neutral molecules, which is seen in figure 25. Neutral molecules show a lower RMSE than anions and cations. This agrees with the findings from SMx[10,11] studies.

For the P3NI, P4 and more accurate models the anions always show higher RMSE than the cations. For the P3 model, 51 setups have higher cation RMSE. They use the $Geom^{PCM}$ or $Geom^{0}$ dataset exclusively in combination with the simulated solvent susceptibility functions. These 51 setups are among the worst RMSE for the complete datasets using the P3 model. The lowest RMSE is 2.55 kcal/mol. There are more than 50 variants of the P3 model with a lower RMSE.

The Minnesota solvation database consists of 112 ions and 31 ionic water clusters. The clusters are composed of one water molecule in the optimal geometry to the ion. These clusters are benchmarks how well a model handles explicitly placed water molecules (see chapter 3.2). This is important for continuum based implicit solvation models since this a way to introduce a solvation shell like structure to the calculations. In RISM this in not needed. The solvation shells are included in the local solvent density fluctuations. Table 15 shows that the RMSE of clustered and bare ions are comparable. This indicates that the partial molar volume correction is able to handle explicitly placed water molecules. However they are not needed for the calculations. EC-RISM proves to be very flexible in the case of neutral molecules. If the interaction between a water molecule and the solvent is studied and the solvent molecule is explicitly included the expected error of the calculation does not increase much.

If the solute is charged the addition of a water molecule can be beneficial especially when the ion is a cation. In most cases an addition of water molecules is not needed. However, the performance of the clustered molecules may be improved. In this work EC-RISM calculations were done with GAFF parameters on the oxygen and hydrogen atoms of the water molecules. Using the MSPC/E parameters should lead to a more consistent parametrization and may improve the accuracy.

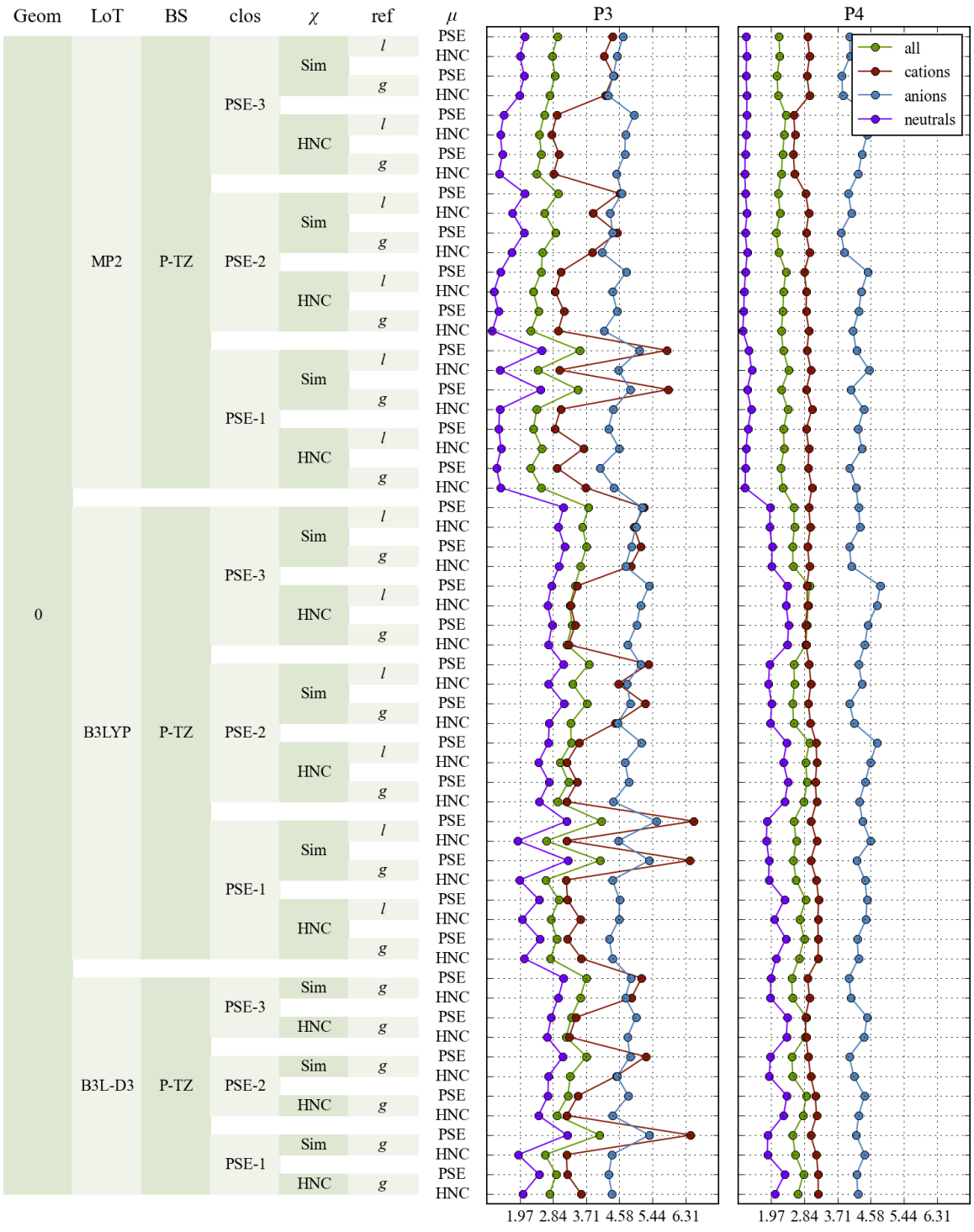


Figure 25: RMSE in kcal/mol of the free energy of solvation of the MNSol. The results are shown for anions, cations, neutral molecules and the complete dataset. Results are shown for the P3 and P4, models with the identifier Model($\text{Geom}_{\text{ref}}^{\text{Geom}}$, QM/LoT/BS, B[clos], χ[χ], F[μ], {EC-RISM})

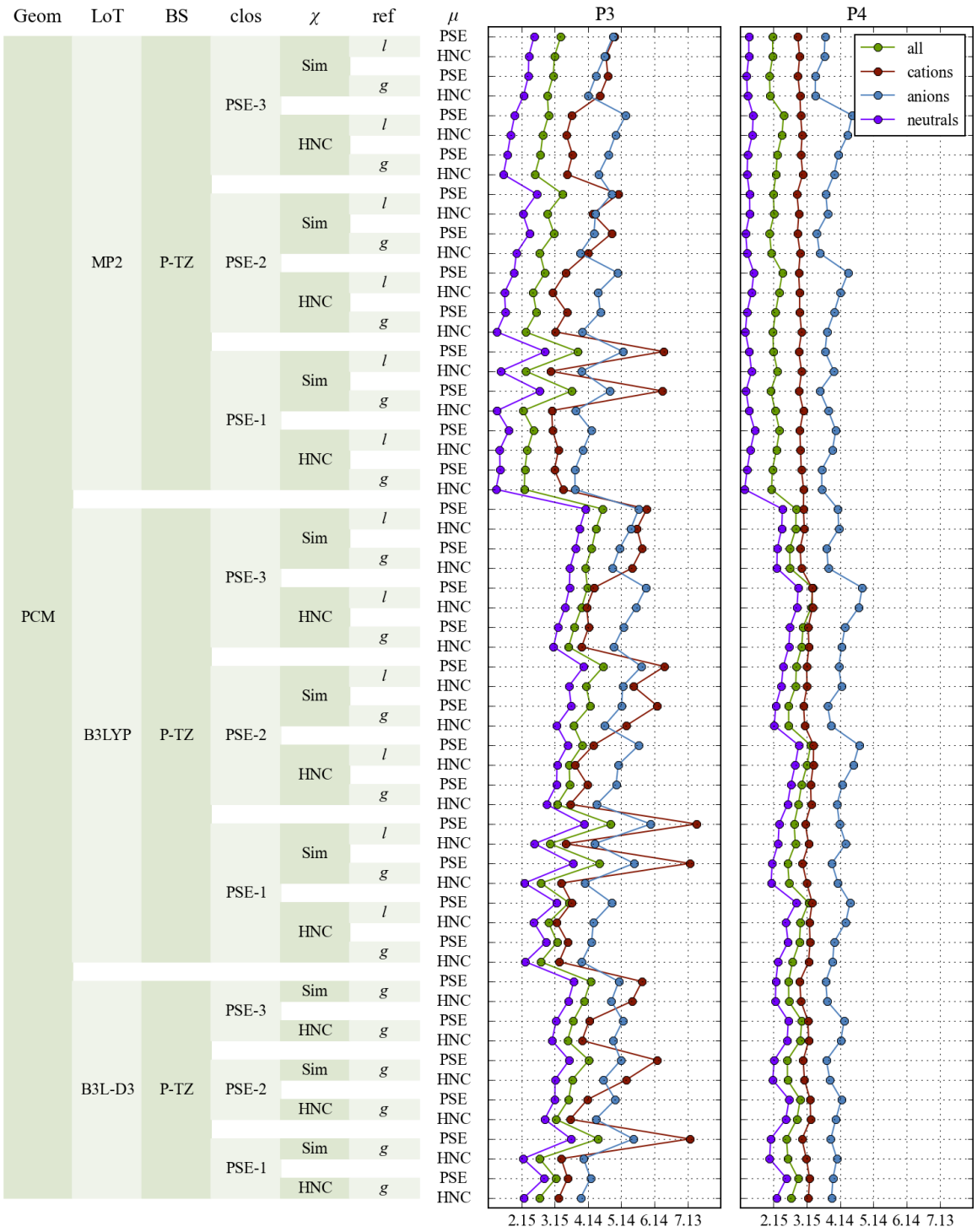


Figure 25 continued

Geom	LoT	BS	clos	χ	ref
PCM	MP2	D-DZ	PSE-3	Sim	*l*
					g
				HNC	*l*
					g
			PSE-2	Sim	*l*
					g
				HNC	*l*
					g
			PSE-1	Sim	*l*
					g
				HNC	*l*
					g
MNSol	MP2	D-DZ	PSE-3	HNC	*l*
DZ	MP2	D-DZ	PSE-3	HNC	*l*
	B3LYP	P-DZ	PSE-2	HNC	*l*
D3	MP2	D-DZ	PSE-3	HNC	*l*
					g
			PSE-2	HNC	*l*
					g
			PSE-1	HNC	*l*
					g
	B3LYP	P-TZ	PSE-3	HNC	*l*
					g
			PSE-1	HNC	*l*
					g
0	MP2	D-DZ	PSE-3	Sim	*l*
					g
				HNC	*l*
					g
			PSE-2	Sim	*l*
					g
				HNC	*l*
					g
			PSE-1	Sim	*l*
					g
				HNC	*l*
					g

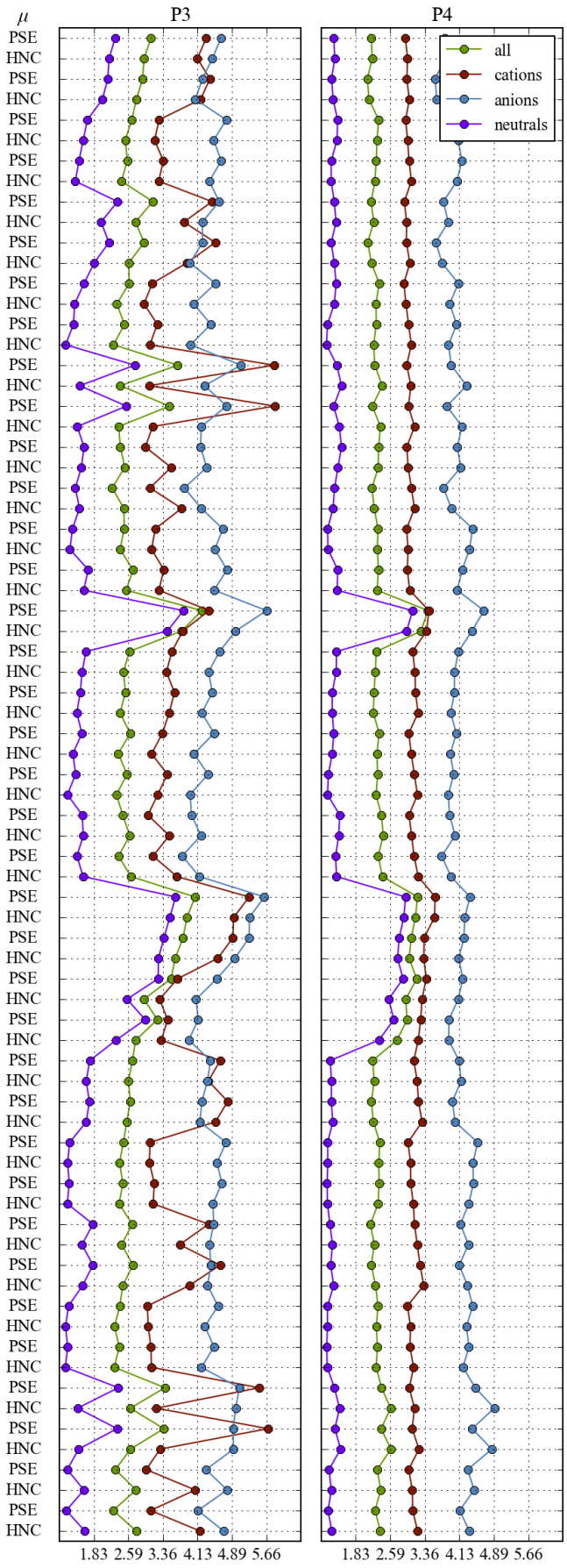


Figure 25 continued

Table 15: RMSE values in kcal/mol of charged molecules in the vicinity of explicitly placed water molecules and bare molecules. The corresponding identifier is el(Geom$_g^{PCM}$, QM/MP2/6-311+G**, *B*[PSE-3], χ[HNC], *F*[HNC], {EC-RISM}).

	all ions		anions		cations	
model	clustered	bare	clustered	bare	clustered	bare
P2	5.20	**4.78**	5.41	**4.88**	**3.45**	4.65
P3	4.58	**4.02**	4.81	**4.32**	**2.44**	3.59
P3NI	3.97	**3.55**	4.05	**3.94**	3.35	**2.98**
P4	3.98	**3.55**	4.07	**3.92**	3.32	**3.00**
P4G	4.56	**4.01**	4.79	**4.32**	**2.45**	3.56
P4PSE	**9.33**	12.45	**5.86**	8.20	20.94	**16.40**
P5	3.64	**3.51**	**3.67**	3.74	3.47	**3.18**
P9	3.18	**3.03**	3.24	**3.11**	**2.70**	2.92

7.3.4 Influence of the setup

Previously the overall performance of the partial molar volume correction was discussed. Now the influence of the settings of the calculations (the theoretical setup) is analyzed. This means parts of the complete dataset (see figure 23) will be focused and discussed under different aspects. Therefore subsets of the complete dataset are analyzed to identify critical points in the EC-RISM setting for the partial molar volume correction.

This chapter is divided into three parts. The partial molar volume correction is based firstly on the QM, secondly on the 3D-RISM (both form the EC-RISM cycle) and thirdly on the way how the correction is applied. In the first part the geometry optimization, level of theory and the basis set are discussed. The 3D-RISM part focusses on the closure and the susceptibility function. Finally the RMSE is compared between the models. The models themselves are influenced by the conformational ensemble, the reference geometry and the excess chemical potential functional. These settings will be discussed.

7.3.4.1 Quantum mechanics setting and optimization

The quantum mechanics were performed in a two-step process. First the geometries were optimized using different levels of theory, basis sets and solvent models. Then the EC-RISM calculations were done at three levels of theory and three different basis sets. The datasets can be separated into two classes. The Geom^0 and Geom^{MNSol} sets were optimized in vacuum, while the other datasets were optimized in PCM solvent.

Figure 26 shows the RMSE for different setups with a focus on the dataset. For the overall dataset, the difference in RMSE introduced by the dataset is usually within 0.1 kcal/mol. Exceptions are low parameter models like the P2 model which show a slight preference of vacuum geometries. Here the RMSE of the vacuum optimized models is about 0.3 kcal/mol smaller. The RMSE of neutral molecules are usually lower for the vacuum optimized structures. Cations are best described by the $Geom^0$ dataset and anions should be optimized using PCM.

Surprisingly the vacuum reference geometries and the molecules using the own geometry as a reference structure have very similar RMSE. Using vacuum optimized geometries as a reference for the solvation free energy is expected to be more accurate because it allows for the treatment of molecule reorganization during solvation. For most molecules in the database, this energetic contribution does not seem to be negligible.

Figure 27 shows the influence of the level of theory by comparing MP2 to B3LYP and B3LYP-D3. In all cases, MP2 has lower RMSE than the density functional based methods. The benefit of MP2 is reduced by increasing the number of parameters in the model. This indicates that the high coefficient models may be slightly over-parametrized, since they overcome deficits of the quantum mechanics calculations. This is dangerous because these deficits may depend on the database diversity quite heavily. In this case the models may not work well on other databases.

The increase in accuracy due to MP2 does also vary with the charge of the molecules. For cations, the P9 model performs almost equally with MP2 and B3LYP. The differences between MP2 and B3LYP are higher for neutral molecules than for ions. This indicates that the ionic solvation free energies may have high experimental errors that are artificially “modeled” (over-fitted) with the partial molar volume correction under certain conditions.

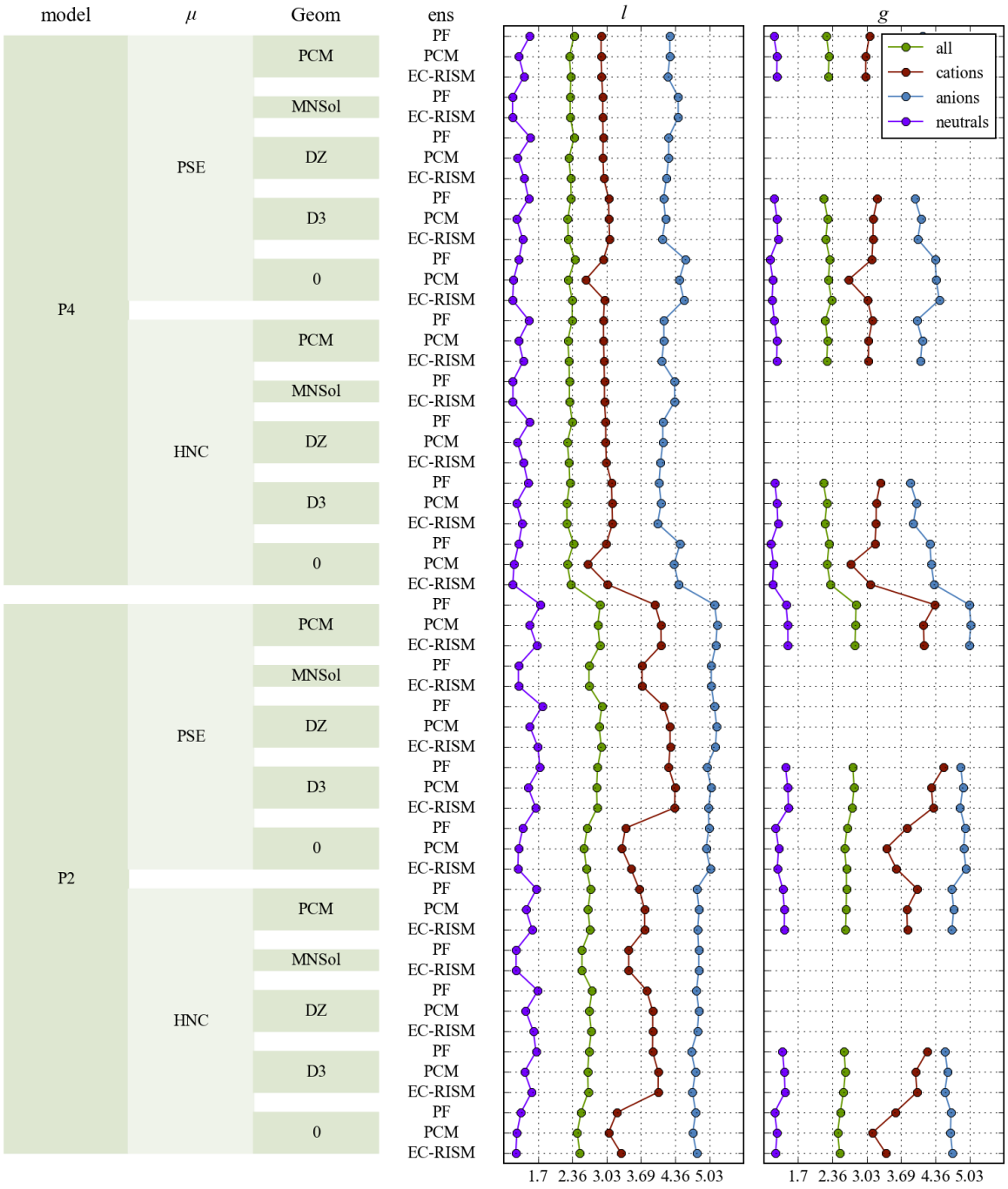


Figure 26: Influence of the dataset. RMSE in kcal/mol of the free energy of solvation of the MNSol calculated for the P2, and P4 models with the identifier Model($\mathrm{Geom}_{\mathrm{ref}}^{\mathrm{Geom}}$, QM/MP2/aug-cc-pVDZ, *B*[PSE-3], χ[HNC], *F*[μ], {ens}).

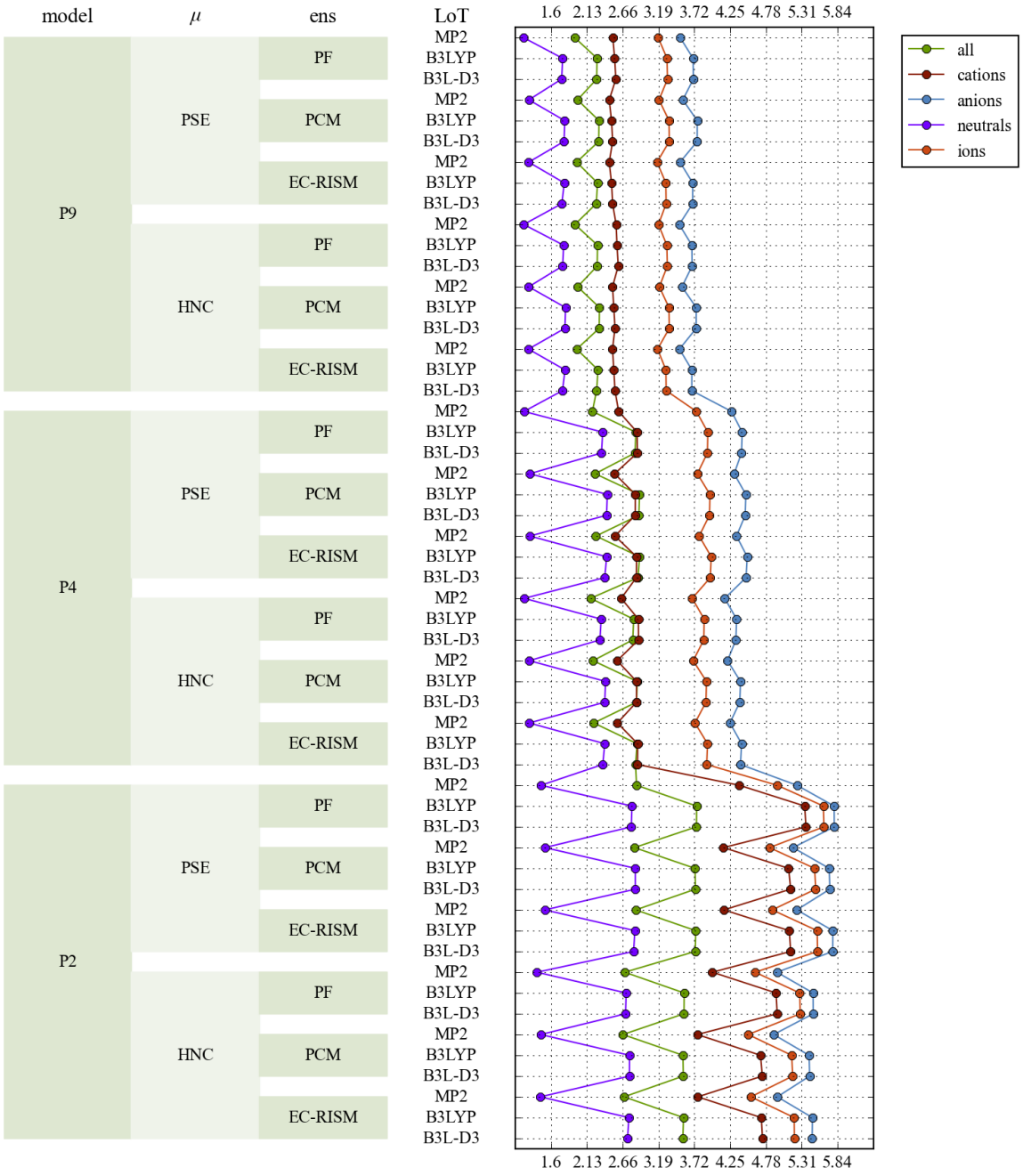


Figure 27: Influence of the level of theory. RMSE in kcal/mol of the free energy of solvation of the MNSol calculated for the P2, P4, and P9 models with the identifier Model(Geom^0_g, QM/LoT/aug-cc-pVDZ, *B*[PSE-3], χ[HNC], *F*[μ], {ens}).

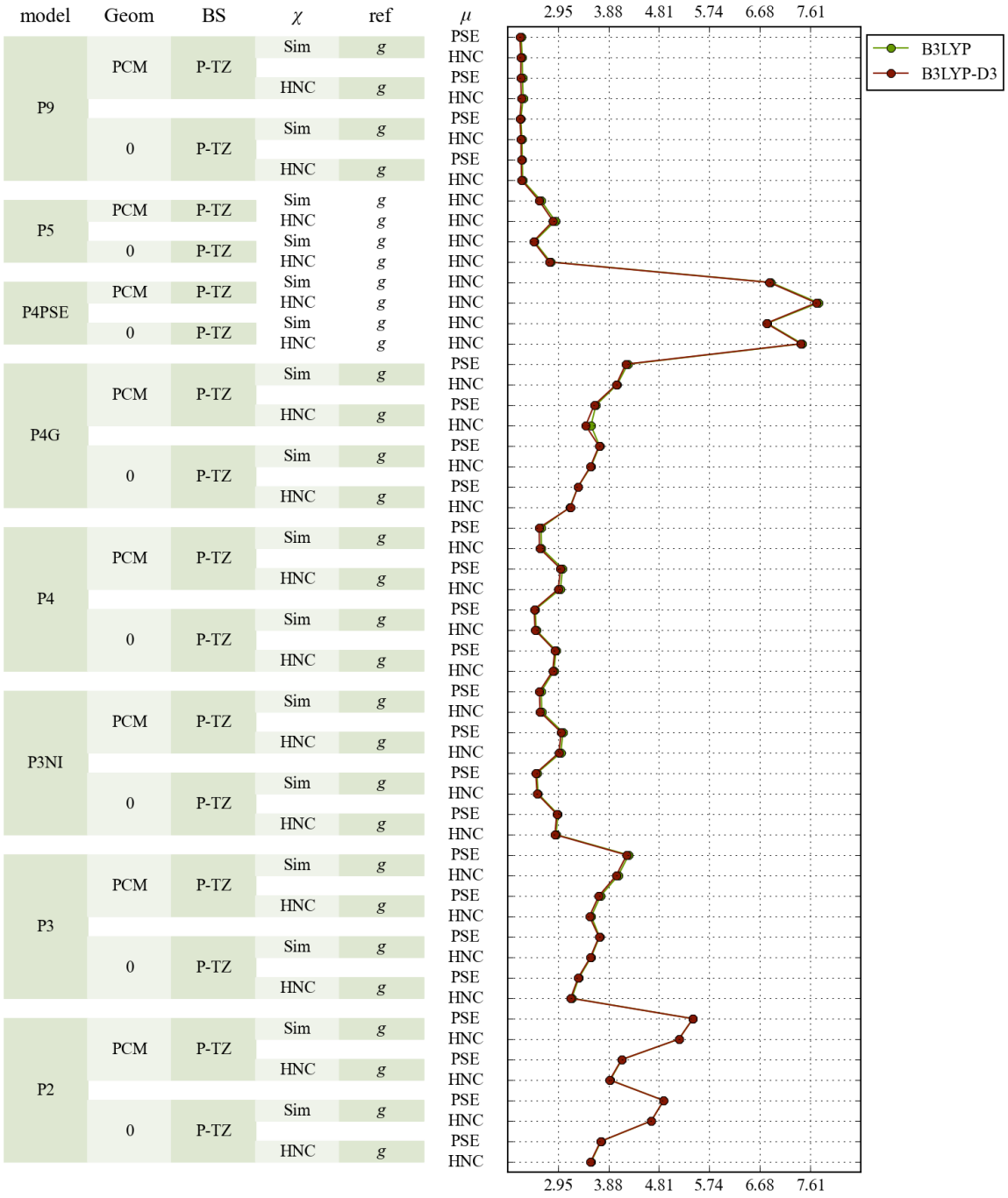


Figure 28: Influence of a D3 correction. RMSE in kcal/mol of the free energy of solvation of the MNSol, calculated for models with the identifier Model($\mathrm{Geom}_{g}^{\mathrm{Geom}}$, QM/LoT/BS, *B*[PSE-3], χ[χ], *F*[μ], {EC-RISM}).

A D3 correction seems to be not beneficial when coupled with the partial molar volume correction as shown in figure 28. The combination of B3LYP-D3 with the $\mathrm{Geom}^{\mathrm{D3}}$ dataset is not possible since the $\mathrm{Geom}^{\mathrm{D3}}$ vacuum reference geometries are not available and the D3 correction cancels in case of $\mathrm{Geom}_{l}^{\mathrm{D3}}$.

When comparing the aug-cc-pVDZ basis set with the 6-311+G** basis set one finds a similar dependency of the error as in case of the level of theory. More parameter models (like P4, P5 or P9) disguise the effect of the basis set. Both basis sets have comparable RMSE for the neutral molecules. As seen in figure 29 the differences are about 0.2 kcal/mol in case of the P2 model and 0.1 kcal/mol for the P4 and P9 models. The aug-cc-pVDZ has lower RMSE than the Pople basis sets.

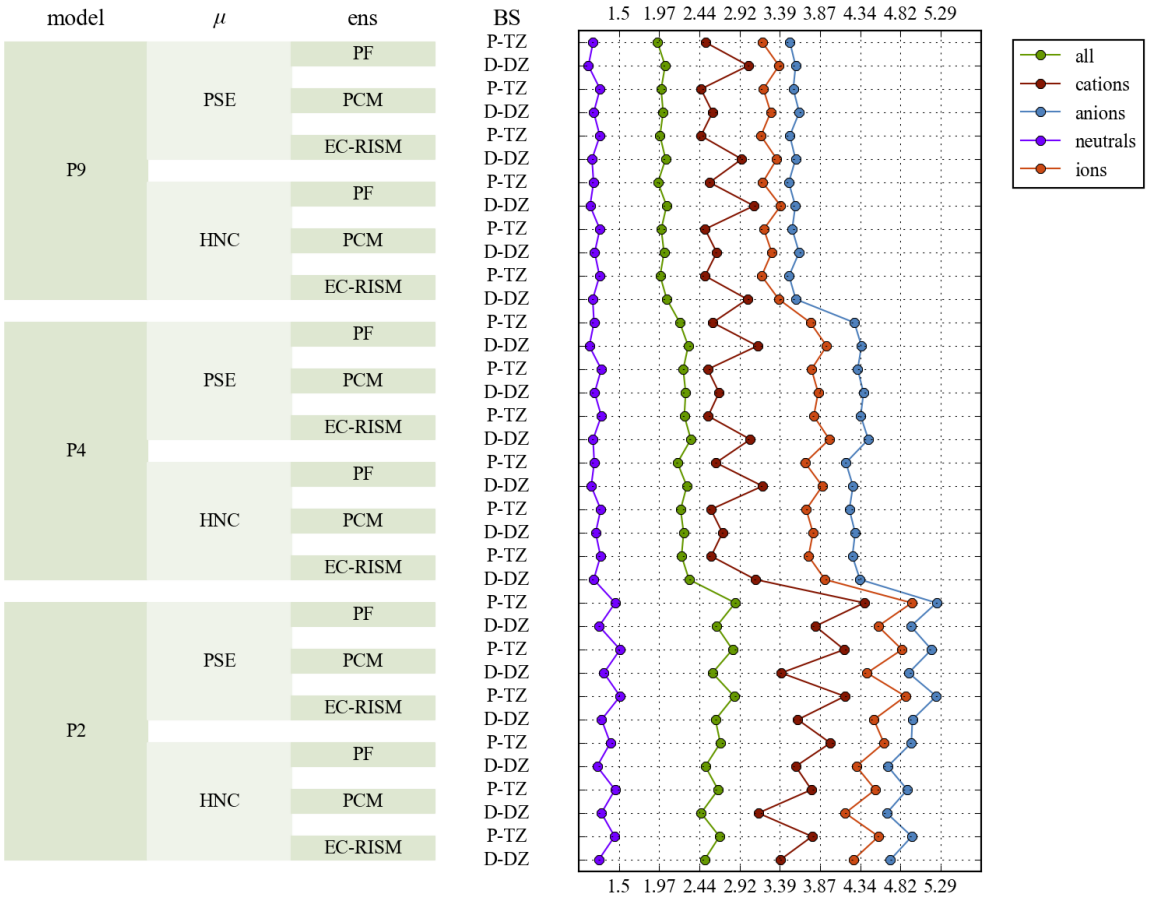


Figure 29: Influence of the basis sets aug-cc-pVDZ and 6-311+G**. RMSE in kcal/mol of the free energy of solvation of the MNSol calculated for the P2, P4, and P9 models with the identifier Model(Geom^0_g, QM/MP2/BS, *B*[PSE-3], χ[HNC], *F*[μ], {ens}).

This changes for ionic molecules. The Dunning basis set still has lower RMSE with the P2 model. However the Pople basis set is better with the P4 and P9 model. When using the aug-cc-pVDZ with ions, the PCM minimum energy structures should be used for the calculations. For the Pople basis set the choice of the minimum energy structure is not as important. For cations using the Pople basis set will decrease the RMSE about 0.4 kcal/mol. For anions this effect is rather small with a change of only about 0.1 kcal/mol.

The aug-cc-pVDZ and 6-311G** basis set are both relatively big and therefore not useable for bigger molecules. Therefore a partial molar volume correction was parameterized with a smaller basis set during optimization and EC-RISM. Also the EC-RISM calculations were done with the PSE-2 closure where convergence is more likely. Using the P9 model the Geom^{DZ} dataset with the 6-31+G* yields a slightly lower RMSD than the Geom^{PCM} dataset with the 6-311+G** basis set. However the difference is smaller than 0.1 kcal/mol. In the case of either the P4 or the P9 model, the Geom^{PCM} dataset with the 6-311+G** has lower RMSE of about 0.2 kcal/mol (figure 30).

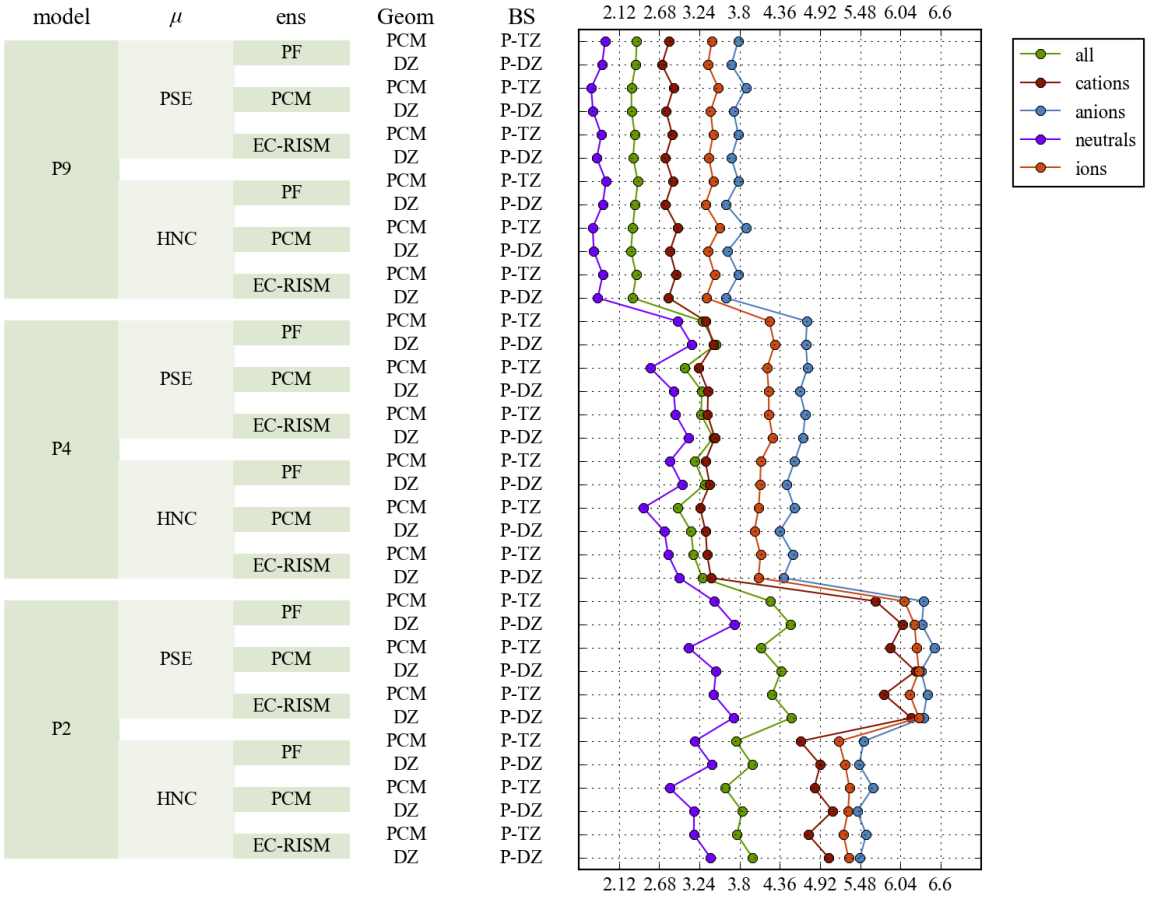


Figure 30: Influence of the basis sets 6-31+G* and 6-311+G**. RMSE in kcal/mol of the free energy of solvation of the MNSol calculated for the P2, P4, and P9 models with the identifier Model(Geom$_l^{DZ}$, QM/B3LYP/BS, *B*[PSE-2], χ[HNC], *F*[μ], {ens}).

7.3.4.2 3D RISM setting

Two RISM related settings were tested in this work. The first is the PSE order, the second is the susceptibility. The PSE order has little influence on the RMSE of neutral molecules. In some cases when the simulated susceptibility function is used, the RMSE calculated with the PSE-1 closure is slightly higher. The PSE-1 closure is the best choice in comparison to the other closures in combination with the P2 and B3LYP/6-311+G** and the HNC functional. In this case a D3 correction does not make a difference.

In the case of the ions, the RMSE differences between the closures are higher. The differences are still small when the P4 or a similar model (P3NI, P5 or P9) is used. However, within the P2 model the choice of the closure is relevant. If the simulated susceptibility functions are used in combination with the PSE functional of the excess chemical potential, then the RMSE calculated with the PSE-1 closure is higher than the RMSE of PSE-2 and PSE-3 closure. If however the HNC functional is used with the simulated solvent susceptibilities then PSE-1 is better than PSE-2 and PSE-3. In this case the RMSE is comparable to the P4 model. When the P2 model is used with the HNC solvent model, an increase in PSE-order will also increase the RMSE. The only exception is the MP2/aug-cc-pVDZ level with HNC functional where the PSE-2 closure is better.

For the P9 model the difference in RMSE for the HNC and simulated susceptibility is about 0.03 kcal/mol for neutral molecules. When the P4 model is used, the difference increases to 0.1 kcal/mol for MP2 and almost 0.4 kcal/mol for B3LYP. When using the P4 model, one may apply the closure with the best convergence behavior for each individual system without a loss in accuracy. At the B3LYP level the simulated solvent susceptibility functions should be used.

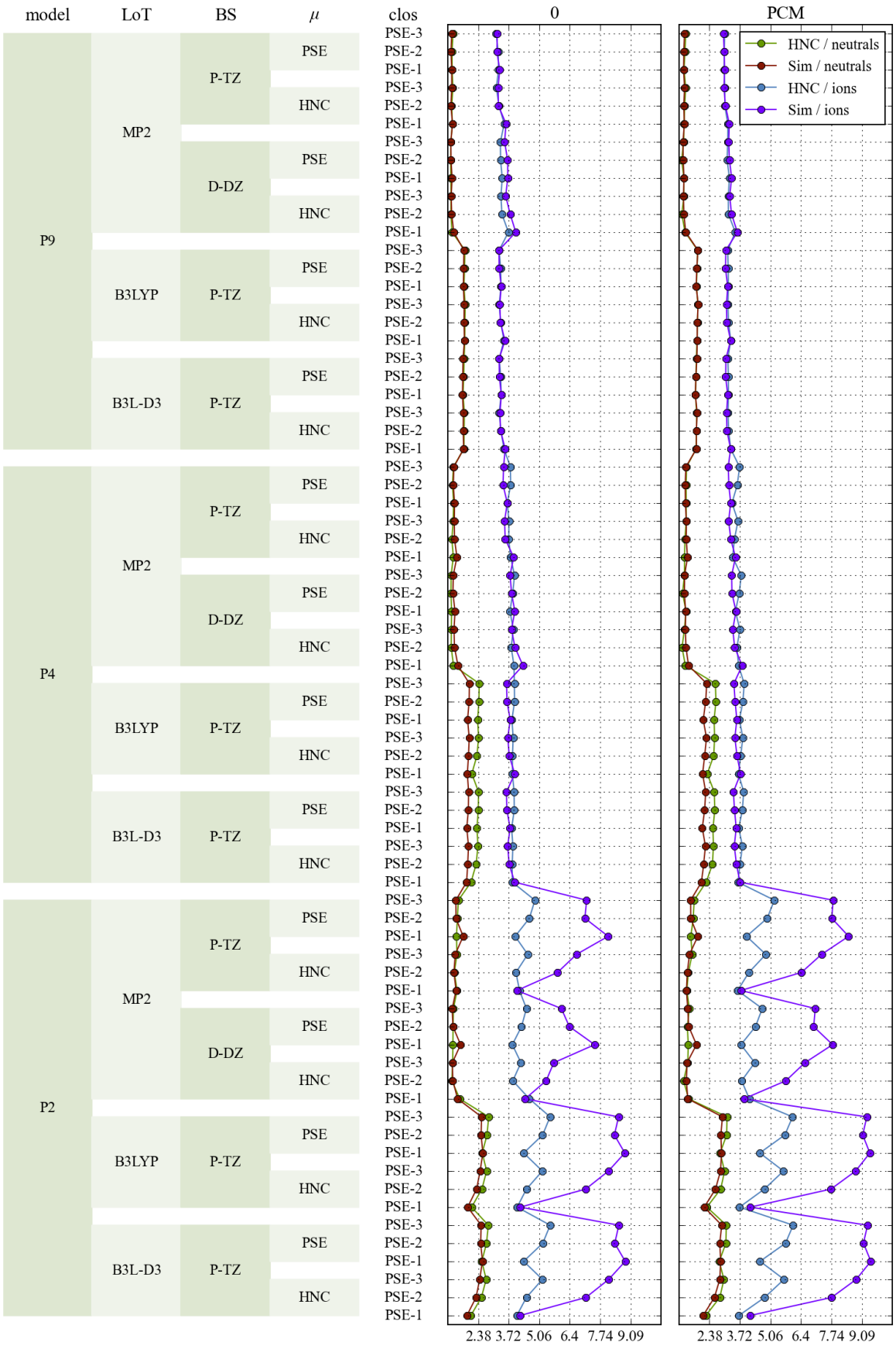


Figure 31: Influence of the closure. RMSE in kcal/mol calculated for models with the identifier Model($\text{Geom}_{\text{g}}^{\text{Geom}}$, QM/LoT/BS, B[PSE-2], $\chi[\chi]$, $F[\mu]$, {EC-RISM}).

7.3.4.3 Partial molar volume correction settings

Figure 32 shows a comparison between the RMSE of all eight models. The left panel shows the models without a c_μ parameter (P2, P3, P4G) together with the P4PSE model which has no c_q parameter. These models have higher RMSE than the other models which are shown in the right panel. Hence these parameters are important for the fitting. From a comparison of the panels the effects of these parameters can be estimated. The difference is highest for Geom^{PCM} dataset when used in combination with the simulated susceptibility function. When the HNC functional is used with the PSE-1 closure, the differences between the P2 model and the P3NI model are small. When the PSE functional is used, the PSE order has only a small influence on the difference between the RMSE. The level of theory has a relatively small impact.

The choice of the chemical potential functional is not as important when the models in the right panels are used as when the models in the left panels are used. For the P3NI model the absolute difference is between 0.003 and 0.243 kcal/mol. For the P3 model the absolute difference may reach 1.8 kcal/mol. With the exception of the P9 model, the HNC functional has lower than average RSME.

The last two parameters in the setup are the reference geometry of the vacuum calculations and the way how the conformational ensemble is averaged. There is hardly any difference between the different ways of ensemble averaging. Usage of the vacuum optimized reference structures lowers the RMSE on the PCM optimized datasets. This is an expected result. The process is closer to the physical process since it includes a structural reorganization of the molecule during the transfer into the solvent. However in the Geom^0 dataset the optimization protocol is similar for the molecule in gas phase and solution. Therefore the RMSE of this dataset is hardly lowered by the usage of vacuum reference geometries. This is shown in figure 33.

The P3NI, P4, P5 and P9 models can be used to correct solvation free energies calculated with EC-RISM. These models have comparably low RMSE. The RMSE of the P3NI, P4 and P5 in close proximity and tend to vary between the setups. The P9 model has the lowest RMSE and the result is less dependent of the setup. The best setups for each model are shown in table 16 and table 17. The RMSE of other well performing setups for each model are given in the results tables in the appendix (chapter 9). In the next chapters further features of the models are discussed and advices are given which model can be used under certain conditions (chapter 7.3.13).

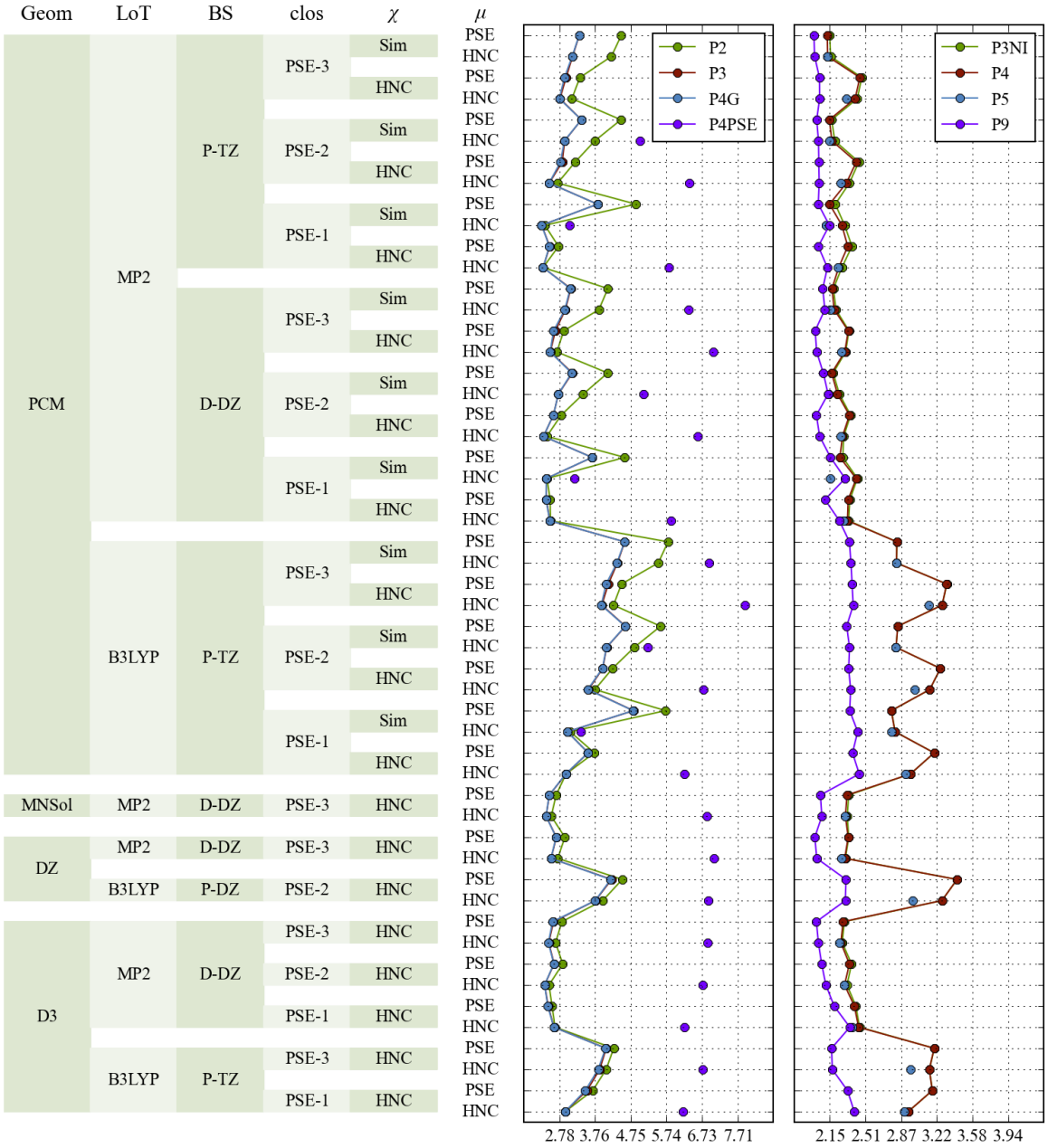


Figure 32: Influence of the model and the excess chemical potential functional. RMSE in kcal/mol of the free energy of solvation of the MNSol calculated all models with the identifier Model($\text{Geom}_l^{\text{Geom}}$, QM/LoT/BS, B[clos], $\chi[\chi]$, $F[\mu]$, {EC-RISM}). The P4PSE and P5 models cannot be evaluated with the PSE functional and therefore every other data point is missing for these models.

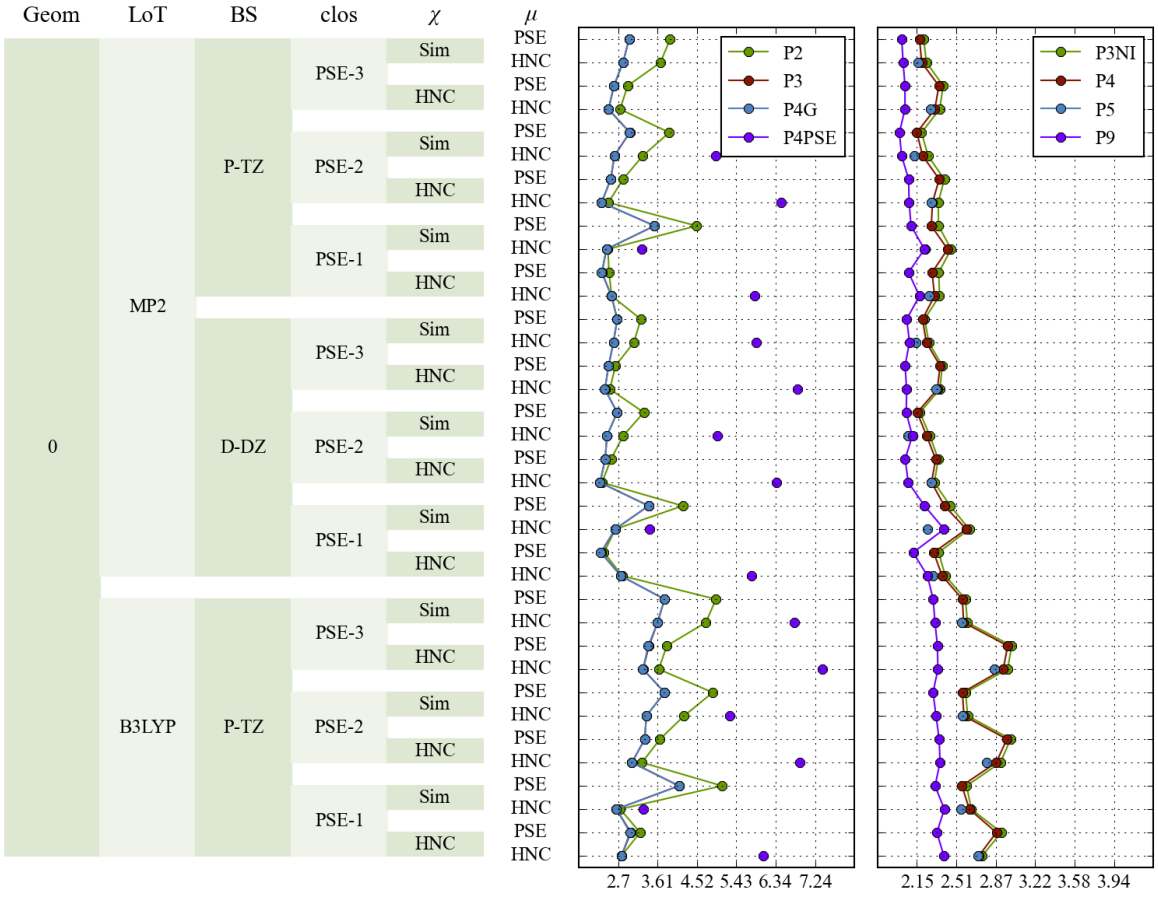


Figure 29 continued

Table 16: Root mean squared errors (kcal/mol) in solvation free energy of the Minnesota solvation database calculated by using different solvent models. For each model the setup with the lowest RMSE is given. All calculations were done with MP2.

model	dataset	basis set	closure	χ	ref	μ	ensemble	all	ions	anions	cations	neutrals
P2	$Geom^{PCM}$	6-311+G**	PSE-1	HNC	*g*	HNC	PF	2.18	3.56	3.68	3.35	1.33
P3	$Geom^{PCM}$	6-311+G**	PSE-1	HNC	*g*	HNC	PF	2.18	3.57	3.67	3.39	1.32
P3NI	$Geom^{PCM}$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	**2.04**	3.26	3.45	2.95	1.32
P4	$Geom^{PCM}$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	**2.0**	3.21	3.37	2.95	1.29
P4G	$Geom^{PCM}$	6-311+G**	PSE-1	HNC	*g*	HNC	PF	2.18	3.56	3.67	3.39	1.32
P4PSE	$Geom^{PCM}$	6-311+G**	PSE-1	Sim	*g*	HNC	EC-RISM	3.0	4.72	5.25	3.68	2.02
P5	$Geom^{PCM}$	6-311+G**	PSE-2	Sim	*g*	HNC	PF	1.97	3.2	3.4	2.84	1.22
P9	$Geom^{PCM}$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	**1.89**	3.03	3.08	2.95	1.21

Table 17: Root mean squared errors (kcal/mol) in solvation free energy of the Minnesota solvation database calculated by using different solvent models. For each model the setup with the lowest RMSE is given. All calculations were done with B3LYP.

model	dataset	basis set	closure	χ	ref	μ	ensemble	all	ions	anions	cations	neutrals
P2	$Geom^{PCM}$	6-311+G**	PSE-1	HNC	*g*	HNC	PF	2.69	3.66	3.9	3.26	2.23
P3	$Geom^{0}$	6-311+G**	PSE-1	Sim	*g*	HNC	EC-RISM	2.65	3.97	4.39	3.18	1.96
P3NI	$Geom^{0}$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	**2.53**	3.68	4.07	2.98	1.96
P4	$Geom^{0}$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	**2.51**	3.62	3.98	2.96	1.96
P4G	$Geom^{0}$	6-311+G**	PSE-1	Sim	*g*	HNC	EC-RISM	2.65	3.97	4.39	3.18	1.96
P4PSE	$Geom^{PCM}$	6-311+G**	PSE-1	Sim	*g*	HNC	EC-RISM	3.2	4.94	5.46	3.95	2.3
P5	$Geom^{0}$	6-311+G**	PSE-3	Sim	*g*	HNC	PF	2.49	3.53	3.91	2.85	1.98
P9	$Geom^{D3}$	6-311+G**	PSE-3	HNC	*l*	PSE	PCM	**2.09**	3.14	3.52	2.33	1.69

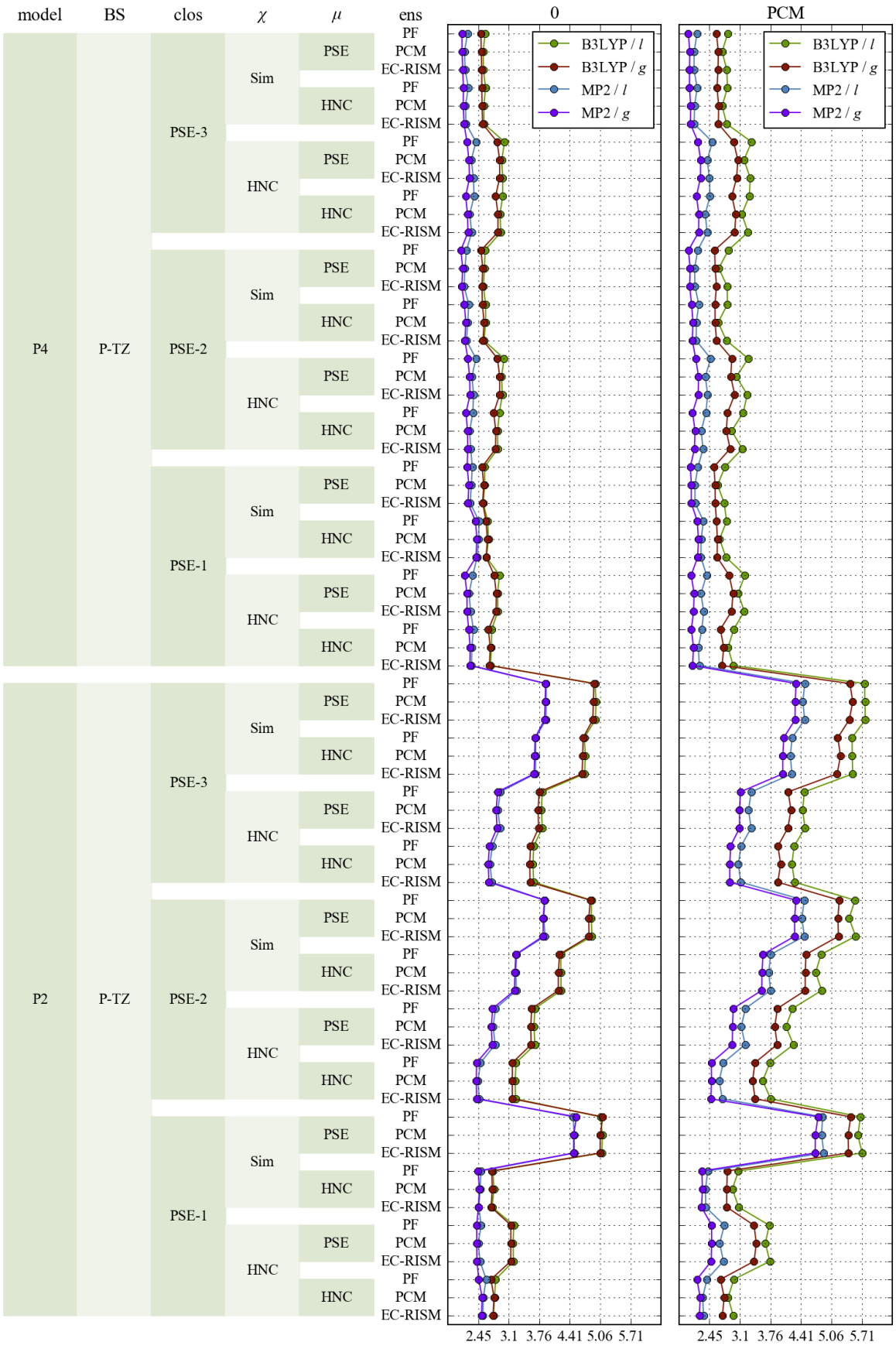


Figure 33: Influence of the reference geometry and ensemble. RMSE in kcal/mol of the models with the identifier Model($\text{Geom}_{\text{ref}}^{\text{Geom}}$, QM/LoT/BS, B[clos], χ[χ], F[μ], {ens}).

7.3.5 Comparison to other solvent models

When comparing the partial molar volume correction to other solvent models, the SM*x* models by Cramer and Truhlar are probably the best references, since these models are parametrized with the same experimental dataset.

In the literature usually the mean unsigned error MUSE is used instead of RMSE. The MUSE is usually smaller than the RMSE but it tends to prefer models with some outliers while an RMSE based parametrization of the model will avoid them. With the SM8 model a MUSE of 1.32 kcal/mol was calculated for the complete database[11]. The subset of ions reached a MUSE of 3.20 kcal/mol. The neutral molecules reached 0.55 kcal/mol. The lowest MUSE of the partial molar volume correction are between 1.45 and 4.78 dependent of the used model. For neutral molecules the SM*x* models yield smaller errors than the partial molar volume correction. However in the partial molar volume correction this error might be reduced by an optimization of the LJ parameters. Here they were taken from the GAFF force field. A tuning of the ε and σ parameters can reduce the error if molecule classes which are discussed in chapter 7.3.10. In contrast to the partial molar volume correction the LJ parameters are needed in the EC-RISM cycle. Such an optimization will need much more computational resources. The partial molar volume correction is more accurate in the description of ions than the SM*x* models.

The number of parameters is an important property of a semi-empirical model. The partial molar volume correction uses between two and nine parameters for the modeling together with the non-bonded parameters of the GAFF force field. The GAFF force field has more than 30 types for the atoms C, N, O, H, S, P, F, Cl, Br and I. Hence there are over 60 parameters (σ_{LJ} and ε_{LJ} for every atom type) used for the description of the solute solvent interactions in this work. They were not optimized. As an example COSMO-RS uses five global parameters and one for every element (independent of any atom type) and SM8 uses multiple per-element parameters and even element-element specific parameters (108 in total).

During this work the GAFF parameters were used without any modifications. Hence the correction has as many adjusted parameters as the used model. It is likely that an optimization of the solvent atom interactions is able to reduce the MUSE of the neutral molecules to a similar value like COSMO-RS or SM8.

Table 18: Mean unsigned errors (kcal/mol) in solvation free energy of the Minnesota solvation database calculated by using different solvent models. The "all" entry of SM8 was calculated from the arithmetic mean of the "neutrals" and "all ions" entries. The number counts were taken from the paper. The identifier is Model($Geom_l^{MNSol}$, QM/ MP2/ aug-cc-pVDZ, *B*[PSE-3], χ[HNC], *F*[HNC], {EC-RISM}). The optimal setup for every model is shown in table 19 and table 20. Please note that COSMO-RS is the only solvent model that was not parameterized with the Minnesota solvation database. Entries that are not reported are labeled with --.

	all	all ions	anions	cations	neutrals
SM8[11]	1.32	3.20	--	--	0.55
SM6[10]	1.86	4.38	--	--	0.54
COSMO-RS[124]	--	--	--	--	0.58
P2	1.70	3.49	3.84	2.93	0.99
P3	1.68	3.25	3.71	2.52	1.06
P3NI	1.60	3.14	3.74	2.19	0.99
P4	1.58	3.13	3.73	2.18	0.96
P4G	1.68	3.25	3.71	2.52	1.06
P4PSE	4.78	9.48	5.97	15.04	2.92
P5	1.58	3.10	3.66	2.21	0.98
P9	1.45	2.68	2.90	2.34	0.96

Table 19: Mean unsigned errors (kcal/mol) in solvation free energy of the Minnesota solvation database calculated by using different solvent models. For each model the setup with the lowest MUSE is given. All calculations were done with MP2.

model	dataset	basis set	closure	χ	ref	μ	ensemble	all	ions	anions	cations	neutrals
P2	$Geom^0$	aug-cc-pVDZ	PSE-1	HNC	*g*	PSE	PF	1.52	2.98	3.19	2.63	0.94
P3	$Geom^0$	aug-cc-pVDZ	PSE-1	HNC	*g*	PSE	PF	1.45	2.89	3.16	2.45	0.88
P3NI	$Geom^{PCM}$	6-311+G**	PSE-2	Sim	*g*	PSE	PF	1.45	2.65	2.83	2.38	1.0
P4	$Geom^{PCM}$	6-311+G**	PSE-2	Sim	*g*	PSE	PF	1.42	2.59	2.73	2.36	0.97
P4G	$Geom^0$	aug-cc-pVDZ	PSE-1	HNC	*g*	PSE	PF	1.45	2.89	3.16	2.45	0.88
P4PSE	$Geom^{PCM}$	6-311+G**	PSE-1	Sim	*g*	HNC	EC-RISM	2.02	3.27	3.64	2.67	1.56
P5	$Geom^{PCM}$	aug-cc-pVDZ	PSE-2	Sim	*g*	HNC	PF	1.36	2.61	2.82	2.27	0.88
P9	$Geom^{PCM}$	6-311+G**	PSE-2	Sim	*g*	PSE	PF	1.33	2.42	2.46	2.35	0.93

Table 20: Mean unsigned errors (kcal/mol) in solvation free energy of the Minnesota solvation database calculated by using different solvent models. For each model the setup with the lowest MUSE is given. All calculations were done with B3LYP.

model	dataset	basis set	closure	χ	ref	μ	ensemble	all	ions	anions	cations	neutrals
P2	$Geom^0$	6-311+G**	PSE-1	HNC	*g*	HNC	PF	2.0	3.14	3.6	2.41	1.58
P3	$Geom^0$	6-311+G**	PSE-1	Sim	*l*	HNC	EC-RISM	1.98	3.22	3.65	2.53	1.54
P3NI	$Geom^0$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	1.87	2.93	3.29	2.34	1.49
P4	$Geom^0$	6-311+G**	PSE-3	Sim	*g*	PSE	PF	1.86	2.88	3.25	2.3	1.5
P4G	$Geom^0$	6-311+G**	PSE-1	Sim	*l*	HNC	EC-RISM	1.98	3.22	3.65	2.53	1.54
P4PSE	$Geom^0$	6-311+G**	PSE-1	Sim	*g*	HNC	PF	2.3	3.93	4.56	2.91	1.71
P5	$Geom^0$	6-311+G**	PSE-3	Sim	*g*	HNC	PF	1.85	2.8	3.17	2.22	1.52
P9	$Geom^{D3}$	6-311+G**	PSE-3	HNC	*l*	PSE	PCM	1.6	2.49	2.79	1.98	1.35

7.3.6 Training and test molecules

A trained model has to be tested against a group of molecules which was not part of the parametrization to allow for an estimate of the expected error. To do this the Minnesota solvation database was divided into a test set to train the parametrization and a test set to measure the quality of the parametrization with an external source. The MNSol was divided into 15 groups of similar compounds. These groups are in shown in table 27. From each group 10 % or at least two of the molecules (rounded up) were randomly chosen and placed in the test set. This was done ten times individually to generate ten sets of test and training molecules.

In figure 34 the RMSE of the complete dataset is compared to the RMSE of the average RMSE of all training and test sets. The RMSE of the full dataset is always within 1.5 times the standard deviation of all the training sets. The difference is always smaller than 5%. The average RMSE of the test sets is within (one time) the standard deviation of the overall RMSE.

These results show that the models are reliable. One can expect that the error of an unknown compound is close to the RMSE of the chosen model.

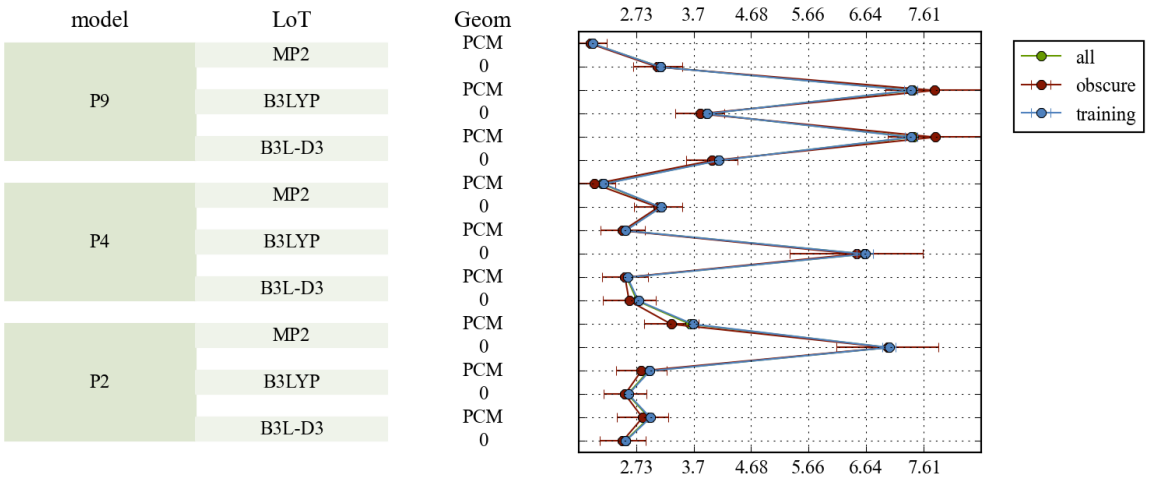


Figure 34: RMSE in kcal/mol for the complete MNSol, the average of all training sets and all test set. The standard deviation for all averages is labelled. The identifier is Model($\text{Geom}_g^{\text{Geom}}$, QM/LoT/6-311+G**, *B*[PSE-3], χ[HNC], *F*[HNC], {EC-RISM}).

7.3.7 Distributions of prediction errors

When building a statistical model the residuals (in the case the error of the molecules) should be normally distributed after the fitting process. Then the error can be assumed to stem from some kind of random noise. To test this, the error quantiles were correlated to the quantiles of a normal distribution and the coefficients of determination were calculated.

The coefficients of determination are between 0.883 and 0.999 for most of the parameterizations. Only the HNC/BF-0 setup shows a lower correlation. Here the coefficients of determination are between 0.747 and 0.992. The low correlations appear here when the P4PSE model is used. This is not surprising, since the PSE functional has no physical meaning in this case.

When the average determination per model is calculated, the P3NI, P4, P5 and P9 models have the highest correlations with values between 0.974 and 0.977. The P3 and P4G models have values of 0.964 and the P2 and P4PSE model have values of 0.954 and 0.956. From this perspective all models are well behaved.

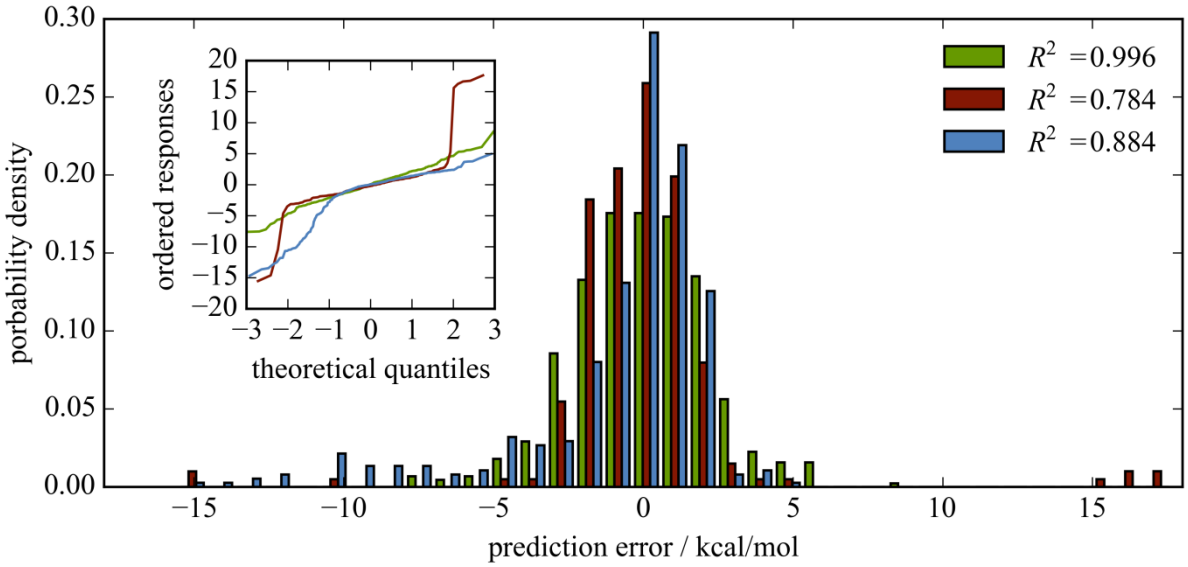


Figure 35: The inset plot shows the error quantiles plotted against the theoretical quantiles of a normal distribution. The main figure shows the corresponding histograms. **Red:** P4PSE($Geom_l^{PCM}$, QM/MP2/aug-cc-pVDZ, *B*[HNC-BF0], χ[HNC-BF0], *F*[HNC], {EC-RISM}); **Green:** P9($Geom_g^{PCM}$, QM/B3LYP/6-311+G**, *B*[PSE-2], χ[MD], *F*[HNC], {EC-RISM}); **Blue:** P2($Geom_g^0$, QM/MP2/aug-cc-pVDZ, *B*[PSE-2], χ[MD], *F*[PSE], {EC-RISM}).

7.3.8 Model parameters

In this chapter the optimized coefficients of the models are discussed. There are two questions to be answered. At first the parts of the setup heavily influencing the parametrization are identified. In a second step the model coefficients are correlated along the different setups. Lists with all coefficients for each model can be found in the supporting information. The units of the coefficients are given in table 29 in the appendix.

The mean standard deviation of the coefficients within a part of the setup gives insight into the importance of this part. When for example the choice of the dataset is important for the

parametrization, the coefficients will not change much between different parametrizations using the same dataset. Hence the standard deviation in each dataset is small in comparison to the overall standard deviation. For the P4 and P9 model this is shown in figure 36. Here only the $Geom^{PCM}$ and $Geom^{0}$ datasets are discussed. They have similar amounts of setups and will not lead to an overestimation of certain parameters.

In case of the two datasets, the internal variations are relatively big. Within all models, there is no coefficient with a mean standard deviation smaller than 80 % of the absolute standard deviation. All coefficients differ much between the setups. The most stable coefficient within the datasets is the c_I coefficient. Therefore c_I changes significantly between the optimization conditions. This can be useful for the calculation of relative free energies. The calculation of free energy changes will be correct, even if the needed setup is not parameterized for the desired dataset, but for a different one. Due to the difference in the calculation of relative free energies, the systematically shifted effect of the c_I coefficient will cancel.

The choice of the ensemble or reference geometry does not introduce variance. Hence the used ensemble and reference geometry have almost no impact on the parametrization. A similar statement can be made about the influence of the chemical potential functional. However, at the low coefficient models the relative standard deviation may drop below 90 % for some coefficients. The stability of the parametrization against changes in the chemical potential is surprising. Since the HNC functional and the PSE-*k* functional lead to different chemical potentials, a bigger effect of this coefficient was expected.

When looking at the QM settings of the EC-RISM calculations, level of theory and basis set have a big impact on the parametrization. Compared to the influence of the level of theory the influence of the basis set is smaller.

The 3D-RISM related settings, closure and susceptibility function both have a big impact on the parametrization. The different closures and susceptibility functions seem to produce very different free energies before the corrections are applied. Hence using a different closure or solvent model as used during the parametrization will be prone to errors.

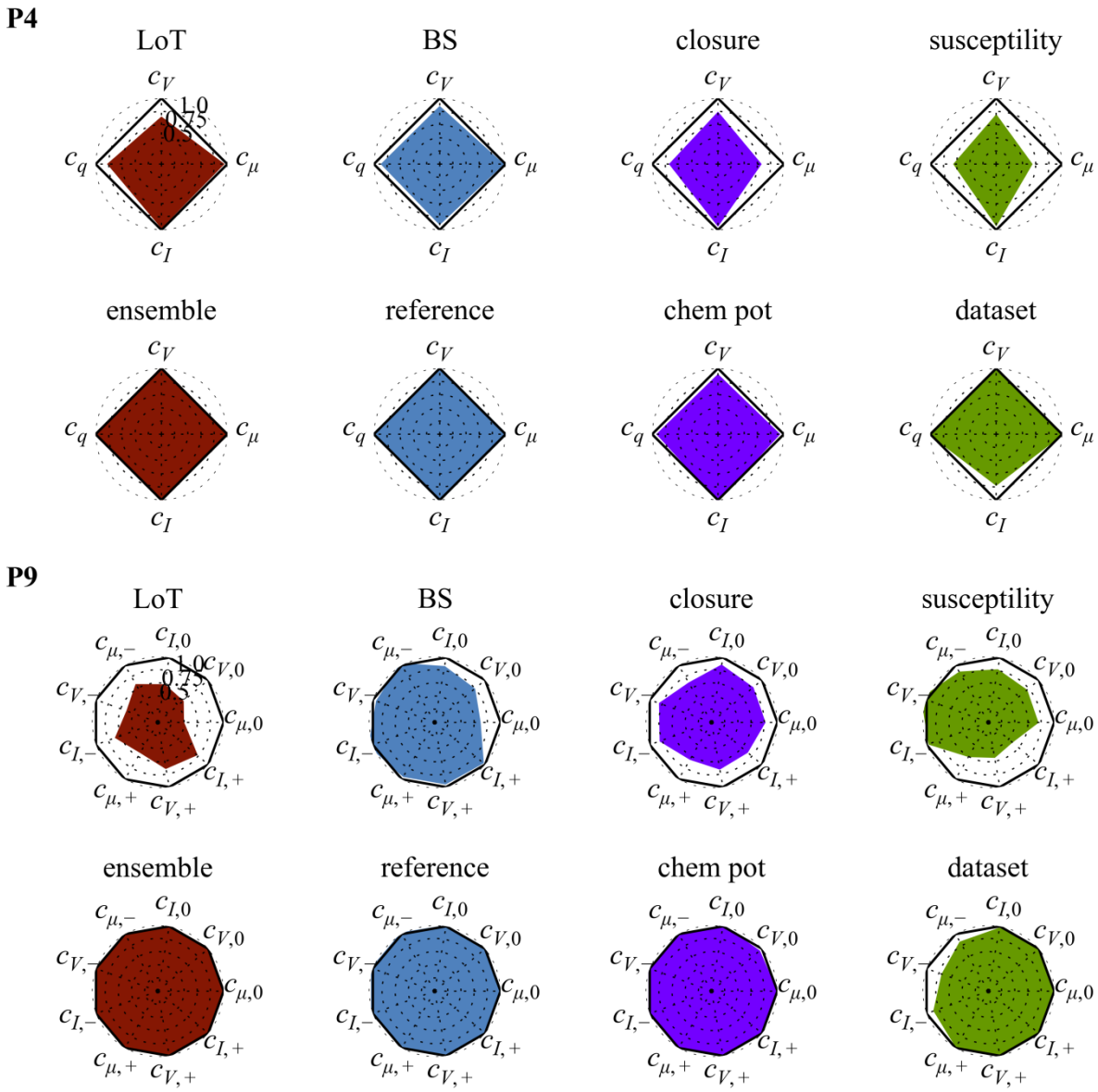


Figure 36: Mean standard deviation of every parameter over all setups within each influence. The standard deviation is relative to the overall standard deviation of all influences. This is labelled by the black line. On top is the P4 model, bottom is the P9 model.

In an ideal model every part transports a unique part of information. In this case the coefficients are independent of each other. In table 21 to table 24 the cross correlation matrices of the parameters were calculated. The majority of the correlation coefficients is independent. Not surprisingly correlation is more likely in models with four and more parameters. Greater correlation coefficient absolute values than 0.7 appear in the P4 model, the P4PSE, the P5 and the P9 model. In the P4 model, there is an anti-correlation between c_V and c_q. In the P4PSE and the P5 model these values are also close to minus one (-0.66 and -0.73). Higher scaling coefficients of the partial molar volume reduce the need of the charge term. Unfortunately, the

coefficient does not increase with the level of theory of the calculations. One will not get rid of charge dependent parameter by increasing the level of theory.

The P4PSE and P5 models have further coefficient correlations. In the P4PSE model there is a strong correlation between c_μ and $c_{d\mu}$. In the P5 model c_μ correlates the c_q. P9 model has nine coefficient correlation coefficients greater than 0.7 (absolute values). Among them are $c_{I,0}$ and $c_{I,+}$ as well as $c_{\mu,0}$ and $c_{\mu,+}$. This is in agreement with earlier findings. The cations are well described by most of the models. So it is expected, that the correction of the cations is similar to the neutrals'. In each group, there is a correlation between the volume coefficient and the intercept. In all three cases this coefficient is greater than 0.8. In this model the intercept coefficients play the role of the charge coefficient in the other models. Therefore this result is comparable to the high correlation between c_V and c_q in the P4 model. Finally all parameters within the anion and cation groups are correlated. In the case of the anions this correlation is really high with absolute coefficients greater than 0.9. From the perspective of coefficient correlation, the P4PSE, P5 and P9 models seem to be over-parametrized.

There are potentially over-parameterized models and some parts of the setup have little influence on the parametrization. This is important for users of the partial molar volume correction. For the sake of readability this important information is discussed in separate chapters. The question what model should be used with which setup is discussed in chapter 7.3.13. What can be done if the needed model is not parameterized is shown in chapter 7.3.9.

Table 21: Coefficient correlations in the P2, P3 and P3NI model.

P2		
	c_V	c_q
c_V	1.00	-0.56
c_q	-0.56	1.00

P3			
	c_V	c_q	c_I
c_V	1.00	-0.27	-0.53
c_q	-0.27	1.00	-0.61
c_I	-0.53	-0.61	1.00

P3NI			
	c_μ	c_V	c_q
c_μ	1.00	-0.46	0.46
c_V	-0.46	1.00	-0.57
c_q	0.46	-0.57	1.00

Table 22: Coefficient correlations in the P4 and P4G model.

P4				
	c_μ	c_V	c_q	c_I
c_μ	1.00	-0.54	0.50	0.24
c_V	-0.54	1.00	-0.45	-0.86
c_q	0.50	-0.45	1.00	0.01
c_I	0.24	-0.86	0.01	1.00

P4G				
	c_G	c_V	c_q	c_I
c_μ	1.00	0.16	-0.36	0.15
c_V	0.16	1.00	-0.28	-0.52
c_q	-0.36	-0.28	1.00	-0.61
c_I	0.15	-0.52	-0.61	1.00

Table 23: Coefficient correlations in the models using the PSE-*k* functional

P4PSE				
	c_μ	$c_{d\mu}$	c_V	c_I
c_μ	1.00	0.78	-0.66	-0.57
$c_{d\mu}$	0.78	1.00	-0.52	-0.42
c_V	-0.66	-0.52	1.00	-0.12
c_I	-0.57	-0.42	-0.12	1.00

P5					
	c_μ	$c_{d\mu}$	c_V	c_q	c_I
c_μ	1.00	-0.57	-0.73	0.82	0.03
$c_{d\mu}$	-0.57	1.00	0.28	-0.22	0.05
c_V	-0.73	0.28	1.00	-0.63	-0.64
c_q	0.82	-0.22	-0.63	1.00	-0.14
c_I	0.03	0.05	-0.64	-0.14	1.00

Table 24: Coefficient correlations in the P9 model

P9									
	$c_{\mu,0}$	$c_{V,0}$	$c_{I,0}$	$c_{\mu,-}$	$c_{V,-}$	$c_{I,-}$	$c_{\mu,+}$	$c_{V,+}$	$c_{I,+}$
$c_{\mu,0}$	1.00	-0.72	0.30	0.31	-0.23	0.15	0.80	-0.63	0.68
$c_{V,0}$	-0.72	1.00	-0.83	-0.25	0.24	-0.14	-0.61	0.87	-0.60
$c_{I,0}$	0.30	-0.83	1.00	0.00	-0.09	0.00	0.24	-0.69	0.29
$c_{\mu,-}$	0.31	-0.25	0.00	1.00	-0.95	0.96	0.65	-0.42	0.50
$c_{V,-}$	-0.23	0.24	-0.09	-0.95	1.00	-0.93	-0.60	0.40	-0.45
$c_{I,-}$	0.15	-0.14	0.00	0.96	-0.93	1.00	0.55	-0.40	0.50
$c_{\mu,+}$	0.80	-0.61	0.24	0.65	-0.60	0.55	1.00	-0.77	0.89
$c_{V,+}$	-0.63	0.87	-0.69	-0.42	0.40	-0.40	-0.77	1.00	-0.86
$c_{I,+}$	0.68	-0.60	0.29	0.50	-0.45	0.50	0.89	-0.86	1.00

7.3.9 Applying suboptimal parametrizations

In this part the result of the partial molar volume correction is calculated on incorrect setups. The change of the RMSE is a measure for the robustness of the correction. Additionally a user

of the correction gets a hint which correction he or she can use when the calculations were not performed with a parameterized model.

Table 25: Mean RMSE (kcal/mol) changes when a suboptimal model is applied. Every model shown here uses vacuum reverence geometries and the EC-RISM ensemble. When not explicitly mentioned, the HNC solvent model was used.

data	parametrization	mean	min	max
PSE-2	PSE-1	1.0	0.09	3.08
PSE-3	PSE-1	1.41	0.15	4.38
B3LYP	MP2	1.52	1.01	2.2
6-311+G**	aug-cc-pVDZ	0.09	0.03	0.15
$Geom^{PCM}$	$Geom^{0}$	0.65	0.04	2.02
Sim	HNC	13.1	0.02	88.47

Table 25 shows how the use of a wrong parameter set can have little up to significant effects on the RMSE. Using the aug-cc-pVDZ basis set instead of the 6-311+G** basis set does not increase the RMSE by much more than 0.1 kcal/mol. However, using the MP2 model with B3LYP data will likely raise the RMSE by more than 1.0 kcal/mol. The closure shows interesting trends. As expected the RMSE changes more from PSE-3 to PSE-1 than from PSE-2 to PSE-1. When using a wrong closure model to the data, a correction with the HNC functional has to be avoided. This is shown in table 26.

Table 26: Mean RMSE changes when a wrong closure is used. The added error is bigger for the HNC functional models.

data	parametrization	mean(μ^{HNC})	mean(μ^{ex})
PSE-2	PSE-1	1.66	0.23
PSE-3	PSE-1	2.35	0.35

The level of theory is important when the $Geom^{0}$ parametrization is used with the $Geom^{PCM}$ dataset. The B3LYP models are more forgiving. The mean RMSE increase is 0.09 kcal/mol for B3LYP while it is 1.22 kcal/mol for MP2.

Applying the wrong solvent susceptibility function model to the data is likely to introduce a high error to the calculations. If the desired parametrization is not available yet and a suboptimal (parameterized for a different setup) model has to be used instead, the QM calculations

should be as accurate as possible. The MP2/aug-cc-pVDZ average error is increased by 0.96 kcal/mol while the MP2/6-311+G** average error is 8.51 kcal/mol and the average error of the B3LYP is 29.83 kcal/mol.

7.3.10 Errors within substance classes

As already seen in the case of the ions, the accuracy of the free energy correction model differs between substance classes. In the following the differences of RMSE between groups are discussed. The EC-RISM calculations were done with MP2/aug-cc-pVDZ/EC-RISM/PSE-3 with the HNC solvent model. The correction was applied using the same geometry as a reference and the HNC functional. The global minimum was selected from the EC-RISM calculations.

The MNSol is divided into 15 groups of molecules. Group one consists of the three molecules hydrogen, ammonia and water. The groups two, three and four are molecules containing hydrogen, carbon, oxygen and nitrogen. The groups five to nine and 15 are halogenated molecules. The groups ten, eleven and twelve have heavier atoms like sulfur, phosphorus and silicon. Group 13 consists of bare ions and in group 14 are water ion clusters.

Figure 37 shows the RMSE of groups two, three and four on the left panel, the groups five, six and seven are placed on the right panel. Nitrogen containing molecules show higher than average errors. Neutral molecules benefit from an optimization under vacuum conditions.

Within group two (compounds containing H, C and O) the correction works best for hydrocarbons, aldehydes and ketones. Better than average RMSE have hydroxyl and bifunctional molecules. The highest errors have the esters. There is a notable difference in the performance of the vacuum datasets ($Geom^0$ and $Geom^{MNSol}$) and the PCM optimized datasets. This is most dominant for the carboxylic acids. Here the RMSE of the vacuum datasets ($Geom^0 \approx 0.43$ kcal/mol and $Geom^{MNSol} \approx 0.33$ kcal/mol) is about 1.3 kcal/mol smaller than in the other cases (RMSE ≈ 1.7 kcal/mol).

In group three (compounds containing H and/or C and N) aromatic nitrogen heterocycles have below average RMSE (< 0.95 kcal/mol), while nitriles (>1.7 kcal/mol) and hydrazines (>2.5 kcal/mol up to 3.29 kcal/cal) have substantially higher RMSE than the average (1.44 kcal/mol to 1.58 kcal/mol).

In group four the amides and urea derivatives have higher RMSE than nitrohydrocarbons and other bifunctional molecules. The difference between these groups is smaller in the vacuum optimized datasets.

When comparing halogenated molecules (groups five to nine) chlorine substituted molecules have low RMSE with the P4 and P9 model. In contrast to the previous classes, fluorine substituted molecules are better described with the PCM optimized datasets. Multihalogenated carbons have relatively low RMSE ranging from 0.72 kcal/mol for the $Geom^{0}$ dataset to 1.11 kcal/mol for the $Geom^{MNSol}$ dataset. The group of halogenated bifunctional compounds, containing at most H, C, N, O, F, Cl and Br shows an interesting trend. In contrast to group two and three, in this case the nitrogen containing molecules have an at least 0.5 kcal/mol lower RMSE than the compounds that have oxygen atoms.

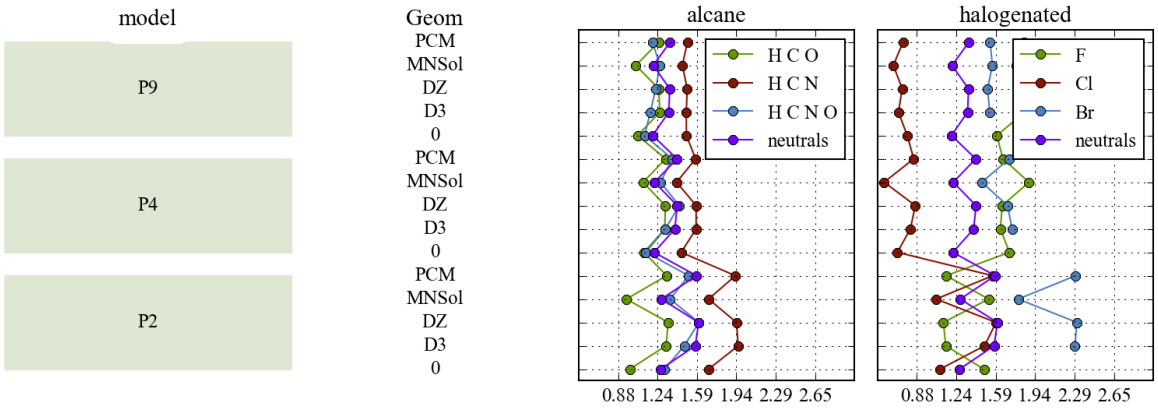


Figure 37: RMSE in kcal/mol for different subclasses of the dataset. The classification corresponds to the classes 2, 3, 4, 5, 6 and 7 of the Minnesota solvation database. The identifier is Model($Geom_{i}^{Geom}$, QM/MP2/aug-cc-pVDZ, *B*[PSE-3], χ[HNC], *F*[HNC], {EC-RISM}).

In figure 38 there are RMSE of molecules with heavier atoms in the left panel and RMSE of ionic molecules in the right panel. The molecules containing sulfur (group ten) have an RMSE profile similar to most of the previously mentioned groups. Again the vacuum geometries are better than the PCM optimized geometries. This is enhanced, when phosphorus is part of a molecule. In this case the RMSE of the vacuum optimized molecules is about one kcal/mol lower than in case of the other datasets.

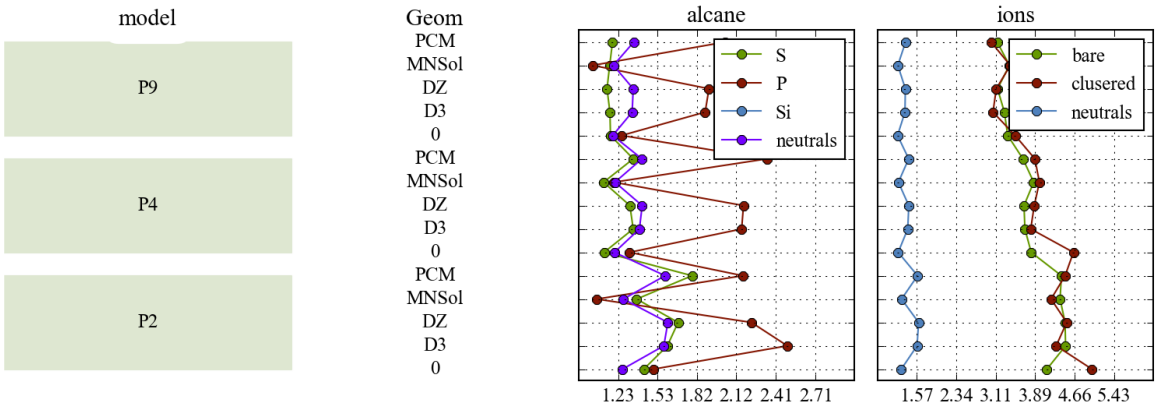


Figure 38: RMSE in kcal/mol for different subclasses of the dataset. The classification corresponds to the classes 10, 11, 12, 13 and 14 of the Minnesota solvation database. Among the clustered molecules in the ion column is also a neutral water-water cluster. The identifier is Model($\text{Geom}_{\text{I}}^{\text{Geom}}$, QM/MP2/aug-cc-pVDZ, *B*[PSE-3], χ[HNC], *F*[HNC], {EC-RISM}).

In the right panel of figure 38 the RMSE of ions is compared to neutral molecules. The differences between clustered and unclustered molecules have been discussed in chapter 7.3.3. Now the focus is on the subclasses of the ionic molecules. The free energy correction works relatively well for molecules with nitrogen and without oxygen. This is similar to the groups of halogenated molecules. Together this indicates that polar molecules with nitrogen are well modelled with the RISM partial molar volume correction, while nitrogen atoms in an apolar environment are not. Another group that is relatively well described, is the group of the halogenated ions which often have below average RMSE. Ions with sulfur are not well described. Here the RMSE is always higher than 5.0 kcal/mol. In contrast to neutral molecules the PCM optimized datasets are better than the vacuum optimized datasets.

7.3.11 Partial molar volume correction with the Verlet closure

The Verlet closure used with the HNC-BF0 formalism has proved to enhance the accuracy of partial molar volumes calculated with 1D RISM[81]. In this chapter the partial molar volume correction is parameterized with the HNC-BF0 closure. Table 11 shows that only for five anions and five cations the EC-RISM iteration was successful. Therefore the ion contribution to the RMSE is smaller in comparison to other models leading to lower RMSE overall. For a meaningful comparison of the Verlet closure with the other models, the RMSE for all models was calculated only for those molecules where iteration converged with the Verlet closure. The results are shown in figure 39. The RMSE calculated with the Verlet closure are slightly better than the results of PSE-2 and PSE-3 when the P3NI or P4 models are used. The P2 and P3 model performs better with the PSE-*k* closures. The P5 and P4PSE models are excluded

since the PSE-functional has no physical meaning when the Verlet closure is used. Due to the small amounts of anions and cations, the P9 model was discarded as well.

The HNC-BF0 closure does not improve the RMSE significantly. The higher computational costs and convergence issues of the HNC-BF0 closure are not justified.

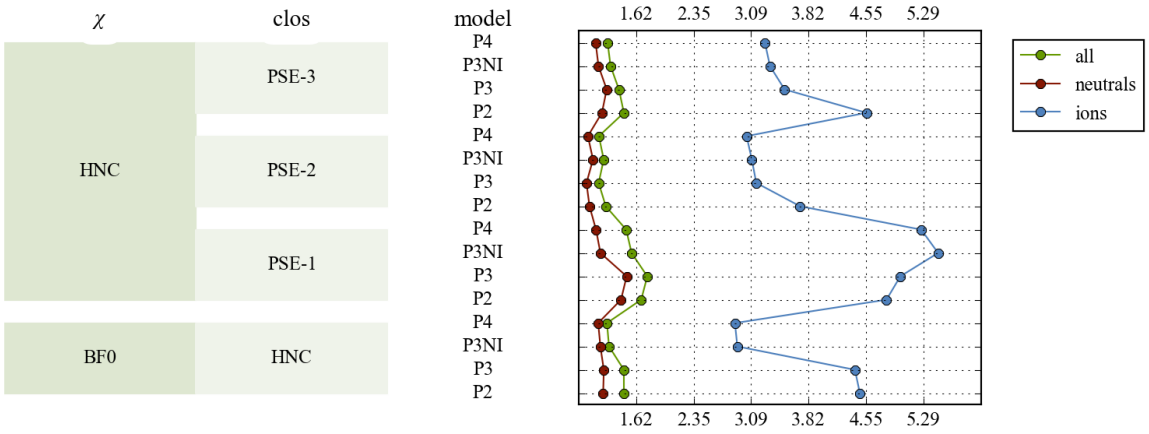


Figure 39: RMSE of the partial molar volume correction with the Verlet closure. The RMSE of the other models was calculated on the reduced subset of molecules which converged in the EC-RSIM calculations. The EC-RISM calculations were performed with the Geom^{PCM} dataset at MP2/6-311+G**/EC-RISM level with the vacuum optimized geometries as a reference. Only the minimum energy structure of the EC-RISM calculations was considered. The free energy was calculated using the HNC functional.

7.3.12 Validation with an external dataset

As an external validation of the parametrization the partial molar volume correction was applied to calculate pK_a values of another dataset. The dataset was compiled by J. Heil from [159] and consists of 39 small molecule pK_a values including anilines, heterocycles, indoles, phenols, pyrroles and thiols.

The molecules in the database were optimized by J. Heil using B3LYP/6-311+G**/PCM. The solvation free energies of these conformations were calculated using MP2/6-311+G**/EC-RISM-PSE-3. Then a partial molar volume correction with the P3NI model was applied using the parametrization for the HNC functional, vacuum reference geometries and the EC-RISM ensemble.

The reaction free energies were calculated from the experimental values using equation (3.11). The theoretical reaction free energies were calculated as shown in chapter 3.2 using the experimental values for the proton solvation free energy and gas phase free energy. The results are shown in figure 40 and table 28 in the appendix.

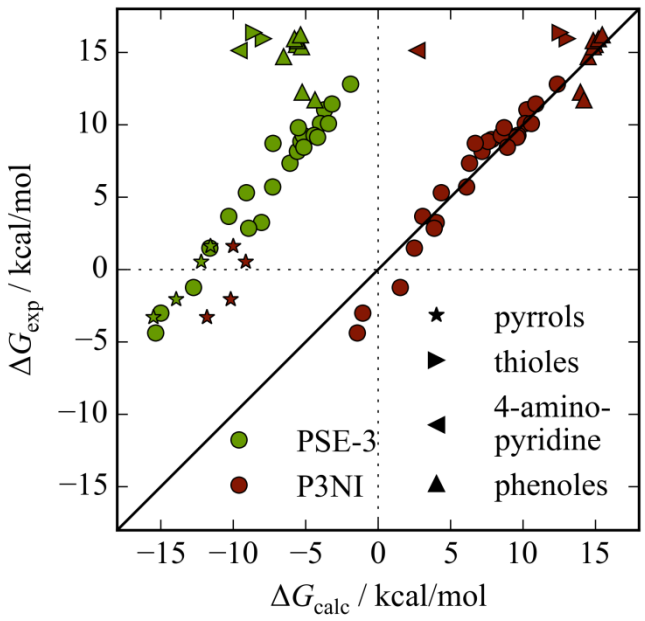


Figure 40: Reaction free energies calculated from experiential pK_a values and the partial molar volume correction. The molecules used here were not part of the parametrization database. Outlaying molecules are labeled with triangles and stars. The green points are PSE-3 free energies (before a partial molar volume correction); the red points have the partial molar volume correction P3NI applied.

The partial molar volume correction reduces the systematic error of the prediction significantly. The correction works well for most molecules. The exceptions are 4-aminopyridine and pyrroles. 4-aminopyridine is an out layer in both cases. The pyrroles are the only group of molecules where the correction worsens the results. Though the ΔG values are slightly closer to the experimental values after the correction, this group is no longer aligned with the other molecule groups as it is the case for the uncorrected PSE-3 data. The partial molar volume correction works well for thioles and phenols. These groups are out layers before the correction. They are in line after the correction is applied. The RMSE for the complete dataset is 3.90 kcal/mol and decreases to 1.20 kcal/mol when the pyrroles and 4-aminopyridine are excluded. The results show that the partial molar volume correction works well for other datasets except the MNSol.

7.3.13 Concluding remarks

The P3NI model has proven to reproduce the free energies of the Minnesota solvation database with a RMSE of 2.11 kcal/mol (see figure 41). It is the model with the best relation between the number of adjustable parameters and minimal RMSE. Therefore it is the recommended model for the calculation of solvation free energies with EC-RISM and a partial mo-

lar volume correction. In chapter 9.3 the setups that work best with the P3NI model are listed. For some exemplarily conditions the identifiers of well performing setups are given.

When B3LYP/6-311+G** is used, the molecules should be prepared with vacuum optimizations. If possible the MD simulation derived susceptibility function should be used. In this case the excess chemical potential functional should be the PSE functional (μ^{PSE}).
P3NI($Geom_g^0$, QM/B3LYP/6-311+G**, *B*[PSE-*k*], χ[MD], *F*[PSE], {PF})

In the case of ions, PSE-2 or PSE-3 in combination with vacuum reference geometries is best.
P3NI($Geom_g^{PCM}$, QM/B3LYP/6-311+G**, *B*[PSE-2], χ[MD], *F*[PSE], {PF})

When only neutral molecules are considered, PSE-1 is the optimal functional and the molecules geometry should be used as reference.
P3NI($Geom_l^0$, QM/B3LYP/6-311+G**, *B*[PSE-1], χ[MD], *F*[PSE], {PF})

When the HNC solvent model is required, PSE-1 and the HNC functional should be used. Usage of vacuum reference geometries and vacuum geometry optimizations are recommended in this case.
P3NI($Geom_g^0$, QM/B3LYP/6-311+G**, *B*[PSE-1], χ[HNC], *F*[HNC], {PF})

When MP2 calculations are feasible, PCM optimized geometries should be used with PSE-2, the simulated susceptibility functions, vacuum reference geometries and the PSE functional for the excess chemical potential.
P3NI($Geom_g^{PCM}$, QM/MP2/6-311+G**, *B*[PSE-2], χ[MD], *F*[PSE], {PF})

Only few settings were tested with the smaller 6-31+G* basis set. For this basis set usage of the HNC functional is recommended.
P3NI($Geom_l^{DZ}$, QM/B3LYP/6-31+G*, *B*[PSE-*k*], χ[HNC], *F*[HNC], {PF})

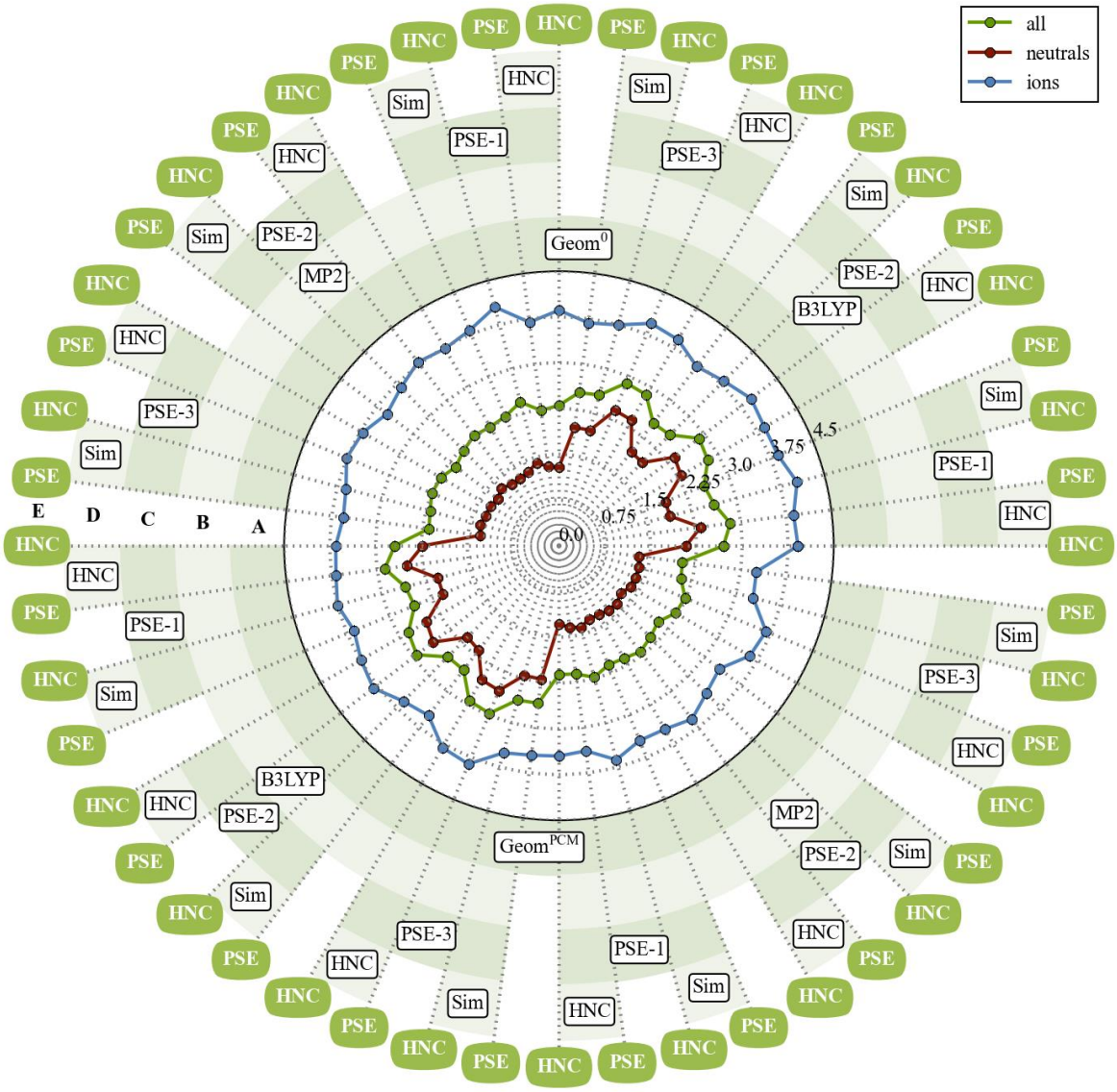


Figure 41: Root mean squared error RMSE in kcal/mol of solvation free energies from the Minnesota solvation database calculated with EC-RISM and the P3NI partial molar volume correction. The inner cycle shows the RMSE and the outer the calculation conditions. The **A** layer describes the QM geometry optimizing in the gas pase with B3LYP/6-311+G** $Geom^0$ and using with B3LYP/6-311+G**/PCM $Geom^{PCM}$. The **B** layer shows the level of theory of the EC-RISM calculations: B3LYP/6-311+G** and MP2/6-311+G**. The **C** layer defines the closure in EC-RISM. The way the solvent susceptibility function was calculated is show in layer **D**. It was calculated using 1D-RISM/HNC or a molecular dynamics simulation. The outer layer **E** shows the functional that was used for the partial molar volume correction. All calculations use vacuum reference geometries and partition functions. P3NI($Geom^A_g$, QM/**B**/6-311+G**, *B*[**C**], χ[**D**], *F*[**E**], {PF})

8 Conclusion

In this work the major goals to rationalize and, possibly, correct systematic prediction errors for free energies calculated within molecular integral equation theories such as RISM were achieved. Starting with fundamental analyses of missing terms in typical approximations for simple fluids a practically usable prediction model for aqueous solvation was developed. This was possible by combining external data (MD simulations and experimental databases) within a properly parametrized RISM formulation.

This work is divided in three parts. In the first two parts it was shown how knowledge from MD simulations can on one hand be used to extract relevant missing information, in particular the so-called bridge function. On the other hand, attempts were formulated and evaluated to make direct use of these data for predicting free energies of simple liquids. The Lennard-Jones (LJ) and the repulsive Weeks-Chandler-Anderson (WCA-rep) potentials for simple fluids were taken as models for these investigations.

The effect of the bridge function on thermodynamic properties calculated with the Ornstein-Zernike equation was studied to increase the accuracy. To do so, the bridge function of the WCA-rep fluid was calculated from MD simulations and compared to the bridge function of the LJ fluid. Regions where bridge functions between the two fluids are similar in certain parametric representations were identified. This knowledge was used to calculate the excess chemical potential for the transition of a WCA-rep particle to a LJ particle with high precision by addition of a bridge dependent functional. The knowledge gained about the properties of those functionals and the numerical stability of their integration turned out to be crucial for the further improvement of the reference interaction site model.

Moving on to practical predictions for realistic solvents, it was demonstrated that the approximate bridge functions from the HNC and PSE-*k* closures can be used to calculate solvation thermodynamics accurately by addition of a correction term. More specifically, the partial molar volume based semi-empirical correction of solvation free energies for more complex systems was successfully integrated with the EC-RISM quantum solvation model. By implementing and testing a vast variety of combinations of quantum-chemical and RISM levels of theory for various forms of correction models, and in contrast to popular continuum models, RISM-predicted hydration free energies now also cover ionic molecules with good accuracy.

In more detail, the semi-empirical correction for hydration free energies calculated with EC-RISM was parameterized also by considering conformational flexibility. Models for charged solutes require at least one parameter for the partial molar volume and another one for the molecular charge. Scaling the RISM excess chemical potential by a third parameter improves the accuracy substantially. With these parameters a root mean squared error (RMSE) of 2.04 kcal/mol can be reached when the correction is parameterized using the entire Minnesota solvation database for aqueous solutions including ions. For neutrals alone, the accuracy is comparable with other state of the art solvent models such as SM8 and COSMO-RS (which are typically highly parametrized to a larger extent than the approach chosen here). However, it outperforms other methods when ionic solvation free energies are to be determined. The correction works well for a verity of QM levels of theory, basis sets and optimization protocols as well as for different 3D RISM closures and solvent susceptibility functions.

Further progress based on optimizing particularly the non-bonded dispersive/repulsive solute-solvent interactions and generalization to other solvents will be a promising approach to extend the scope of integral equation applications in chemistry.

9 Appendix

9.1 Distance geometry in chemical modelling

Distance geometry describes systems with known distances between members, while their actual positions are unknown. In chemistry this can be applied to the generation of conformers for a molecule. The distances in molecules are defined by atom radii, bond lengths, bond angels and other topological properties of the molecule. This information can be taken from libraries of crystal- and NMR structures. When the geometry of a molecule is guessed based on these distances the following steps can be performed[160]:

9.1.1 Identify distance constrains

As an initial step the highest and lowest possible distance between two atoms is specified. For atoms connected by a chemical bond this is straight forward using knowledge from crystal and NMR structures. Minimum and maximum distance specifications can be very complicated when the atoms are not connected directly. In this case they are mostly calculated using the triangle inequality. This results in poorly approximated distance limits. Though this is the most important step of the workflow, it is also the least well solved step[161].

9.1.2 Embedding

In the next step a random distance matrix is generated that fulfills all constrains found in chapter 9.1.1. Since different matrices will lead to different conformations, this step defines the conformation of the molecule. The quality of the sampling of the conformational space is therefore dependent on the algorithm that generates the distance matrix.

The distance matrix is built by setting the value of a randomly chosen distance to a randomly chosen value between the upper and lower limit. The random distance is biased towards the lower and upper limits to enhance the sampling. Than the distance limits are recalculated as in chapter 9.1.1 with the one distance (and the lower and upper limit) set to the random value. This is repeated until all distances in the system have a value assigned to them.

When the distance matrix is know the functional

$$\sum_i \sum_{j<i} \left(\left| \mathbf{x}_i - \mathbf{x}_j \right|^2 - d_{ij}^2 \right)^2 . \tag{9.1}$$

is minimized using standard linear algebra. Here *i* and *j* point at the atoms, **x** is the position of the atom and *d* is the entry of the distance matrix. Usually the minimization is performed in a way that the center of masses of the molecule is placed in the origin.

9.1.3 Refinement

Due to the sum of squares nature of equation (9.1), it is not guaranteed that all constrains are fulfilled in an embedded geometry. Additionally there may be information about the molecule like chirality, bond or torsion angels that may was not utilized in steps 9.1.1 and 9.1.2. This information can be used to clean up the geometry. Often an energy minimization similar to MD simulations is performed.

9.2 Additional tables to chapter 7

Table 27: Composition of training and test sets

name	number	training set	test set
compounds containing H and/or C and O	133	119	14
compounds containing H and/or C and N	45	40	5
compounds containing H, C, N, and O	43	38	5
compounds containing H,C, and F	8	6	2
compounds containing H,C, and Cl	31	27	4
compounds containing H,C, and Br	14	12	2
multihalogen hydrocarbons	12	10	2
Halogenated bifunctional compounds, containing at most H, C, N, O, F, Cl, Br	36	32	4
compounds containing S, but not P	29	26	3
phosphorus compounds	25	22	3
iodine compounds	9	7	2
kations	52	46	6
anions	60	54	6
kation Solute-water clusters	8	6	2
anion Solute-water clusters	23	20	3

Table 28: Protonation free energy $\Delta_r G^{\#}$ of different molecules in kcal/mol. The experimental values were taken from[159]. The values in the PSE-3 and P3NI columns were calculated using MP2/6-311+G**/EC-RISM-PSE-3. The PSE-3 values are uncorrected the P3NI values were corrected with the P3NI model of the partial molar volume correction.

molecule	experimental	PSE-3	P3NI
4-chloroaniline	7.35	-6.08	6.31
4-methoxyaniline	8.99	-5.06	7.83
4-nitroaniline	3.26	-8.06	4.01
aniline	8.17	-5.57	7.16
p-toluidine	8.85	-5.34	7.59
2-aminopyridine	11.03	-3.69	10.26
2-aminothiazole	9.26	-5.18	8.49
2-methylimidazole	12.81	-1.91	12.38
3-aminopyridine	10.08	-3.97	10.15
4-aminopyridine	15.13	-9.6	2.69
4-methylpyridine	10.08	-3.44	10.57
benzimidazole	9.81	-5.51	8.7
imidazole	11.44	-3.18	10.88
isoquinoline	9.26	-4.43	9.64
melamine	8.72	-7.25	6.71
pyrazine	2.85	-8.93	3.88
pyrazole	5.3	-9.09	4.37
pyridine	9.12	-4.19	9.59
pyrimidine	3.67	-10.31	3.07
quinoline	8.44	-5.13	8.91
thiazole	5.71	-7.28	6.12
1-methylindole	-1.24	-12.73	1.52
2-methylindole	1.48	-11.6	2.49
3-methylindole	-4.38	-15.34	-1.45
indole	-3.02	-14.99	-1.07

Table 28 continued

molecule	experimental	PSE-3	P3NI
4-aminophenol	14.72	-6.53	14.49
4-chlorophenol	15.4	-5.29	14.85
4-fluorophenol	15.81	-5.57	14.84
4-hydroxybenzaldehyde	12.26	-5.24	13.95
4-methoxyphenol	15.95	-5.76	15.19
4-methylphenol	16.22	-5.35	15.45
4-nitrophenol	11.72	-4.36	14.21
phenol	15.54	-5.65	14.98
1-me-pyrrol	-2.06	-13.92	-10.17
2-me-pyrrol	1.62	-11.56	-10.0
3-me-pyrrol	0.53	-12.2	-9.14
pyrrol	-3.29	-15.5	-11.81
ethanethiol	16.36	-8.57	12.56
methanethiol	15.95	-7.91	13.04

Table 29: Units of the confidents of the partial molar volume correction models. The elementary charge is labelled *e*. -- means that the coefficient is dimension less.

name	unit
c_V	$\frac{\text{kcal}}{\text{mol Å}^3}$
c_q	$\frac{\text{kcal}}{\text{mol } e}$
c_I	$\frac{\text{kcal}}{\text{mol}}$
c_μ	--
c_G	--
$c_{d\mu}$	--

9.3 Lowest RMSE to the P3NI model

Result 1: RMSE in kcal/mol calculated with the P3NI model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	vac	PSE	PF	2.53	2.98	4.07	1.96
0	PSE-2	Sim	vac	PSE	PF	2.54	3.02	4.08	1.94
0	PSE-3	Sim	vac	PSE	PCM	2.55	2.73	4.15	2.0
0	PSE-3	Sim	vac	PSE	EC-RISM	2.56	2.95	4.1	1.99
0	PSE-3	Sim	vac	HNC	PF	2.56	3.04	4.14	1.96

Result 2: RMSE in kcal/mol calculated with the P3NI model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	self	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	self	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	self	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	self	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	self	HNC	PF	2.68	3.18	4.68	1.88

Result 3: RMSE in kcal/mol calculated with the P3NI model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.69	3.25	3.9	2.23
0	PSE-1	HNC	vac	HNC	PF	2.69	3.26	4.27	2.08
0	PSE-1	HNC	vac	HNC	EC-RISM	2.72	3.22	4.34	2.11
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.73	3.21	3.95	2.28
D3	PSE-1	HNC	vac	HNC	PF	2.74	3.26	3.88	2.33

Result 4: RMSE in kcal/mol calculated with the P3NI model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	self	HNC	EC-RISM	2.74	3.21	4.53	2.06
0	PSE-1	HNC	self	HNC	PCM	2.76	3.22	4.58	2.08
0	PSE-1	HNC	vac	HNC	PF	2.69	3.26	4.27	2.08
0	PSE-1	HNC	vac	HNC	EC-RISM	2.72	3.22	4.34	2.11
0	PSE-1	HNC	self	HNC	PF	2.78	3.19	4.56	2.12

Result 5: RMSE in kcal/mol calculated with the P3NI model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	vac	PSE	PF	2.52	2.99	4.07	1.95
0	PSE-2	Sim	vac	PSE	PF	2.53	3.02	4.08	1.93
0	PSE-3	Sim	vac	PSE	PCM	2.53	2.73	4.15	1.98
0	PSE-3	Sim	vac	PSE	EC-RISM	2.54	2.95	4.1	1.96
0	PSE-2	Sim	vac	PSE	EC-RISM	2.54	2.97	4.11	1.95

Result 6: RMSE in kcal/mol calculated with the P3NI model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	vac	HNC	EC-RISM	2.62	3.18	4.51	1.85
0	PSE-1	Sim	vac	HNC	PF	2.62	3.23	4.47	1.87
0	PSE-1	Sim	vac	PSE	PF	2.56	3.09	4.3	1.87
0	PSE-1	Sim	vac	PSE	EC-RISM	2.57	3.04	4.33	1.88
0	PSE-1	Sim	vac	HNC	PCM	2.66	3.2	4.56	1.9

Result 7: RMSE in kcal/mol calculated with the P3NI model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.67	3.25	3.88	2.21
0	PSE-1	HNC	vac	HNC	PF	2.68	3.26	4.27	2.07
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.69	3.2	3.93	2.23
0	PSE-1	HNC	vac	HNC	EC-RISM	2.7	3.21	4.33	2.08
0	PSE-1	HNC	vac	HNC	PCM	2.73	3.21	4.37	2.12

Result 8: RMSE in kcal/mol calculated with the P3NI model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	vac	HNC	PF	2.68	3.26	4.27	2.07
0	PSE-1	HNC	vac	HNC	EC-RISM	2.7	3.21	4.33	2.08
0	PSE-1	HNC	vac	HNC	PCM	2.73	3.21	4.37	2.12
PCM	PSE-1	HNC	vac	HNC	PF	2.67	3.25	3.88	2.21
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.69	3.2	3.93	2.23

Result 9: RMSE in kcal/mol calculated with the P3NI model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-3	Sim	vac	PSE	PF	2.04	2.95	3.45	1.32
PCM	PSE-2	Sim	vac	PSE	PF	2.04	2.96	3.52	1.27
PCM	PSE-2	Sim	vac	PSE	EC-RISM	2.07	2.89	3.56	1.33
PCM	PSE-3	Sim	vac	PSE	PCM	2.07	2.9	3.51	1.36
PCM	PSE-3	Sim	vac	PSE	EC-RISM	2.07	2.89	3.49	1.37

Result 10: RMSE in kcal/mol calculated with the P3NI model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-3	Sim	vac	PSE	EC-RISM	2.11	2.97	3.65	1.29
PCM	PSE-3	Sim	vac	PSE	PF	2.11	3.05	3.62	1.29
PCM	PSE-2	Sim	vac	PSE	PF	2.11	3.04	3.64	1.26
PCM	PSE-2	Sim	vac	PSE	EC-RISM	2.12	2.97	3.68	1.28
PCM	PSE-3	Sim	vac	PSE	PCM	2.13	2.98	3.69	1.3

Result 11: RMSE in kcal/mol calculated with the P3NI model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	Sim	vac	PSE	PF	2.13	2.97	3.87	1.26
0	PSE-2	HNC	vac	PSE	PF	2.28	2.93	4.22	1.27
PCM	PSE-2	Sim	vac	PSE	PF	2.04	2.96	3.52	1.27
0	PSE-2	HNC	vac	HNC	PF	2.25	3.01	4.1	1.27
PCM	PSE-1	HNC	vac	HNC	PF	2.11	3.11	3.64	1.28

Result 12: RMSE in kcal/mol calculated with the P3NI model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	vac	PSE	PF	2.26	3.06	3.98	1.18
PCM	PSE-2	HNC	vac	HNC	PF	2.21	3.12	3.81	1.18
0	PSE-2	HNC	vac	PSE	PF	2.3	3.09	4.28	1.19
D3	PSE-2	HNC	vac	HNC	PF	2.27	3.26	3.85	1.2
D3	PSE-2	HNC	vac	PSE	PF	2.3	3.18	3.95	1.2

Result 13: RMSE in kcal/mol calculated with the P3NI model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.11	3.11	3.64	1.28
PCM	PSE-1	HNC	vac	PSE	PF	2.12	3.04	3.59	1.35
PCM	PSE-2	HNC	vac	HNC	PF	2.13	3.04	3.71	1.31
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.14	3.04	3.7	1.32
PCM	PSE-1	HNC	vac	HNC	PCM	2.16	3.04	3.74	1.34

Result 14: RMSE in kcal/mol calculated with the P3NI model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	PSE	PF	2.18	3.13	3.74	1.34
D3	PSE-3	HNC	vac	HNC	PF	2.2	3.27	3.89	1.27
D3	PSE-3	HNC	vac	PSE	PF	2.21	3.21	3.97	1.26
PCM	PSE-2	HNC	vac	HNC	PF	2.21	3.12	3.81	1.18
PCM	PSE-1	HNC	vac	PSE	EC-RISM	2.21	3.07	3.81	1.39

Result 15: RMSE in kcal/mol calculated with the P3NI model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	vac	PSE	PF	2.28	2.93	4.22	1.27
0	PSE-2	HNC	vac	HNC	PF	2.25	3.01	4.1	1.27
PCM	PSE-1	HNC	vac	HNC	PF	2.11	3.11	3.64	1.28
0	PSE-3	HNC	vac	PSE	PF	2.25	2.58	4.3	1.29
0	PSE-1	HNC	vac	HNC	PF	2.3	3.09	4.25	1.29

Result 16: RMSE in kcal/mol calculated with the P3NI model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	vac	HNC	PF	2.21	3.12	3.81	1.18
PCM	PSE-2	HNC	vac	PSE	PF	2.26	3.06	3.98	1.18
0	PSE-2	HNC	vac	PSE	PF	2.3	3.09	4.28	1.19
0	PSE-3	HNC	vac	PSE	PF	2.33	3.1	4.36	1.2
D3	PSE-2	HNC	vac	HNC	PF	2.27	3.26	3.85	1.2

Result 17: RMSE in kcal/mol calculated with the P3NI model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	self	HNC	PCM	3.12	3.33	4.35	2.75
DZ	PSE-2	HNC	self	HNC	EC-RISM	3.27	3.4	4.4	2.96
DZ	PSE-2	HNC	self	PSE	PCM	3.27	3.36	4.63	2.88
DZ	PSE-2	HNC	self	HNC	PF	3.31	3.38	4.45	2.99
DZ	PSE-2	HNC	self	PSE	EC-RISM	3.43	3.45	4.67	3.09

Result 18: RMSE in kcal/mol calculated with the P3NI model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	self	HNC	PCM	3.12	3.33	4.35	2.75
DZ	PSE-2	HNC	self	PSE	PCM	3.27	3.36	4.63	2.88
DZ	PSE-2	HNC	self	HNC	EC-RISM	3.27	3.4	4.4	2.96
DZ	PSE-2	HNC	self	HNC	PF	3.31	3.38	4.45	2.99
DZ	PSE-2	HNC	self	PSE	EC-RISM	3.43	3.45	4.67	3.09

9.4 Lowest RMSE to the P2 model

Result 19: RMSE in kcal/mol calculated with the P2 model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.69	3.26	3.9	2.23
0	PSE-1	HNC	vac	HNC	PF	2.72	3.49	4.31	2.07
0	PSE-1	Sim	vac	HNC	EC-RISM	2.73	3.49	4.63	1.92
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.73	3.26	3.96	2.27
D3	PSE-1	HNC	vac	HNC	PF	2.74	3.26	3.88	2.33

Result 20: RMSE in kcal/mol calculated with the P2 model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	self	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	self	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	self	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	self	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	self	HNC	PF	2.68	3.18	4.68	1.88

Result 21: RMSE in kcal/mol calculated with the P2 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.69	3.26	3.9	2.23
0	PSE-1	HNC	vac	HNC	PF	2.72	3.49	4.31	2.07
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.73	3.26	3.96	2.27
D3	PSE-1	HNC	vac	HNC	PF	2.74	3.26	3.88	2.33
D3	PSE-1	HNC	self	HNC	PCM	2.76	3.25	4.08	2.3

Result 22: RMSE in kcal/mol calculated with the P2 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	self	HNC	EC-RISM	2.78	3.5	4.57	2.05
0	PSE-1	HNC	self	HNC	PCM	2.8	3.5	4.63	2.06
0	PSE-1	HNC	vac	HNC	PF	2.72	3.49	4.31	2.07
0	PSE-1	HNC	vac	HNC	EC-RISM	2.77	3.55	4.4	2.09
0	PSE-1	HNC	self	HNC	PF	2.82	3.55	4.61	2.1

Result 23: RMSE in kcal/mol calculated with the P2 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	vac	HNC	PF	2.67	3.25	3.88	2.21
PCM	PSE-1	HNC	vac	HNC	EC-RISM	2.69	3.24	3.93	2.22
0	PSE-1	Sim	vac	HNC	EC-RISM	2.71	3.5	4.63	1.88
0	PSE-1	HNC	vac	HNC	PF	2.71	3.48	4.31	2.05
0	PSE-1	Sim	vac	HNC	PF	2.74	3.63	4.62	1.9

Result 24: RMSE in kcal/mol calculated with the P2 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	vac	HNC	EC-RISM	2.71	3.5	4.63	1.88
0	PSE-1	Sim	vac	HNC	PF	2.74	3.63	4.62	1.9
0	PSE-1	Sim	vac	HNC	PCM	2.75	3.52	4.68	1.93
0	PSE-1	HNC	vac	HNC	PF	2.71	3.48	4.31	2.05
0	PSE-1	HNC	vac	HNC	EC-RISM	2.74	3.53	4.39	2.06

Result 25: RMSE in kcal/mol calculated with the P2 model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	3.88	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.24	3.93	2.22
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.48	4.31	2.05
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.53	4.39	2.06
PCM	PSE-1	HNC	*g*	HNC	PCM	2.75	3.24	3.96	2.25

Result 26: RMSE in kcal/mol calculated with the P2 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.48	4.31	2.05
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.53	4.39	2.06
0	PSE-1	HNC	*g*	HNC	PCM	2.78	3.53	4.43	2.1
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	3.88	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.24	3.93	2.22

Result 27: RMSE in kcal/mol calculated with the P2 model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.35	3.68	1.33
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.24	3.38	3.76	1.39
PCM	PSE-1	HNC	*g*	HNC	PCM	2.26	3.39	3.79	1.4
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	2.28	3.3	4.02	1.37
PCM	PSE-1	Sim	*g*	HNC	PF	2.29	3.43	4.01	1.35

Result 28: RMSE in kcal/mol calculated with the P2 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	PSE	PF	2.31	3.21	4.23	1.24
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.32	3.1	4.31	1.22
0	PSE-2	HNC	*g*	HNC	PCM	2.32	3.1	4.28	1.26
0	PSE-2	HNC	*l*	HNC	PCM	2.32	3.07	4.32	1.25
0	PSE-2	HNC	*g*	HNC	PF	2.32	3.21	4.26	1.21

Result 29: RMSE in kcal/mol calculated with the P2 model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.42	3.42	4.42	1.28
0	PSE-2	Sim	*g*	HNC	PF	3.26	6.45	5.59	1.29
0	PSE-2	Sim	*g*	HNC	EC-RISM	3.23	6.3	5.59	1.31
0	PSE-2	Sim	*l*	HNC	PCM	3.26	6.38	5.73	1.31
0	PSE-2	HNC	*l*	HNC	PCM	2.45	3.22	4.58	1.32

Result 30: RMSE in kcal/mol calculated with the P2 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.35	3.38	4.08	1.21
0	PSE-2	HNC	*g*	HNC	PF	2.32	3.21	4.26	1.21
0	PSE-3	Sim	*g*	PSE	PF	3.19	7.99	5.18	1.22
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.32	3.1	4.31	1.22
0	PSE-3	Sim	*l*	HNC	EC-RISM	3.06	7.0	5.23	1.23

Result 31: RMSE in kcal/mol calculated with the P2 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.35	3.68	1.33
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.24	3.38	3.76	1.39
PCM	PSE-1	HNC	*g*	HNC	PCM	2.26	3.39	3.79	1.4
PCM	PSE-1	HNC	*l*	HNC	PCM	2.31	3.28	4.07	1.4
PCM	PSE-1	HNC	*l*	HNC	EC-RISM	2.32	3.24	4.0	1.5

Result 32: RMSE in kcal/mol calculated with the P2 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	PSE	PF	2.31	3.21	4.23	1.24
0	PSE-1	HNC	*g*	PSE	EC-RISM	2.32	3.1	4.28	1.27
0	PSE-2	HNC	*l*	HNC	PCM	2.32	3.07	4.32	1.25
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.32	3.1	4.31	1.22
0	PSE-2	HNC	*g*	HNC	PF	2.32	3.21	4.26	1.21

Result 33: RMSE in kcal/mol calculated with the P2 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.42	3.42	4.42	1.28
0	PSE-2	HNC	*l*	HNC	PCM	2.45	3.22	4.58	1.32
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.42	3.25	4.44	1.33
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.35	3.68	1.33
0	PSE-2	HNC	*g*	HNC	PCM	2.41	3.25	4.4	1.34

Result 34: RMSE in kcal/mol calculated with the P2 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.35	3.38	4.08	1.21
0	PSE-2	HNC	*g*	HNC	PF	2.32	3.21	4.26	1.21
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.32	3.1	4.31	1.22
0	PSE-2	HNC	*l*	HNC	EC-RISM	2.33	3.04	4.39	1.23
PCM	PSE-2	HNC	*g*	HNC	PCM	2.37	3.22	4.15	1.24

Result 35: RMSE in kcal/mol calculated with the P2 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.83	5.09	5.43	3.16
DZ	PSE-2	HNC	*l*	HNC	PF	3.97	4.92	5.46	3.41
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.98	5.04	5.47	3.39
DZ	PSE-2	HNC	*l*	PSE	PCM	4.37	6.24	6.32	3.47
DZ	PSE-2	HNC	*l*	PSE	PF	4.5	6.06	6.33	3.72

Result 36: RMSE in kcal/mol calculated with the P2 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.83	5.09	5.43	3.16
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.98	5.04	5.47	3.39
DZ	PSE-2	HNC	*l*	HNC	PF	3.97	4.92	5.46	3.41
DZ	PSE-2	HNC	*l*	PSE	PCM	4.37	6.24	6.32	3.47
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	4.52	6.19	6.35	3.71

9.5 Lowest RMSE to the P3 model

Result 37: RMSE in kcal/mol calculated with the P3 model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.65	3.18	4.39	1.96
0	PSE-1	Sim	*g*	HNC	PF	2.66	3.27	4.36	1.98
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.66	3.19	4.54	1.91
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.45	1.98
0	PSE-1	Sim	*l*	HNC	PF	2.67	3.15	4.56	1.94

Result 38: RMSE in kcal/mol calculated with the P3 model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 39: RMSE in kcal/mol calculated with the P3 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.69	3.28	3.89	2.23
0	PSE-1	HNC	*g*	HNC	PF	2.72	3.53	4.31	2.06
D3	PSE-1	HNC	*g*	HNC	PF	2.73	3.3	3.85	2.32
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.73	3.28	3.96	2.27
D3	PSE-1	HNC	*g*	HNC	PCM	2.75	3.31	3.92	2.32

Result 40: RMSE in kcal/mol calculated with the P3 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.78	3.55	4.57	2.03
0	PSE-1	HNC	*l*	HNC	PCM	2.8	3.55	4.62	2.05
0	PSE-1	HNC	*g*	HNC	PF	2.72	3.53	4.31	2.06
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.77	3.58	4.39	2.08
0	PSE-1	HNC	*l*	HNC	PF	2.82	3.6	4.6	2.09

Result 41: RMSE in kcal/mol calculated with the P3 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.63	3.19	4.39	1.93
0	PSE-1	Sim	*g*	HNC	PF	2.65	3.28	4.36	1.96
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.44	1.97
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.22

Result 42: RMSE in kcal/mol calculated with the P3 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.63	3.19	4.39	1.93
0	PSE-1	Sim	*g*	HNC	PF	2.65	3.28	4.36	1.96
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.44	1.97
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05

Result 43: RMSE in kcal/mol calculated with the P3 model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.22
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05
PCM	PSE-1	HNC	*g*	HNC	PCM	2.74	3.27	3.95	2.25

Result 44: RMSE in kcal/mol calculated with the P3 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05
0	PSE-1	HNC	*g*	HNC	PCM	2.78	3.58	4.43	2.09
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.22

Result 45: RMSE in kcal/mol calculated with the P3 model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.39	3.67	1.32
PCM	PSE-1	Sim	*g*	HNC	PF	2.2	3.16	3.74	1.4
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	2.2	3.06	3.78	1.42
PCM	PSE-1	Sim	*g*	HNC	PCM	2.22	3.06	3.83	1.43
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.29

Result 46: RMSE in kcal/mol calculated with the P3 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	2.22	3.18	3.78	1.37
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.23	3.07	3.83	1.41
PCM	PSE-1	HNC	*g*	PSE	PCM	2.25	3.07	3.87	1.42
0	PSE-1	HNC	*g*	PSE	EC-RISM	2.25	3.08	4.14	1.21

Result 47: RMSE in kcal/mol calculated with the P3 model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-2	HNC	*g*	HNC	PCM	2.25	2.98	4.12	1.25
0	PSE-2	HNC	*l*	HNC	PCM	2.3	2.94	4.33	1.25
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.26	2.97	4.17	1.25
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.29

Result 48: RMSE in kcal/mol calculated with the P3 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.25	3.18	3.93	1.16
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
0	PSE-2	HNC	*g*	HNC	PF	2.26	3.19	4.14	1.17
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.28	3.1	4.21	1.19
PCM	PSE-2	HNC	*g*	HNC	PCM	2.28	3.07	4.02	1.2

Result 49: RMSE in kcal/mol calculated with the P3 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.39	3.67	1.32
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.29
PCM	PSE-1	HNC	*g*	PSE	PF	2.23	3.28	3.71	1.45
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-1	HNC	*g*	PSE	PCM	2.24	2.94	4.03	1.36

Result 50: RMSE in kcal/mol calculated with the P3 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	2.22	3.18	3.78	1.37
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.23	3.07	3.83	1.41
0	PSE-1	HNC	*g*	PSE	EC-RISM	2.25	3.08	4.14	1.21
0	PSE-1	HNC	*g*	PSE	PCM	2.25	3.09	4.11	1.23

Result 51: RMSE in kcal/mol calculated with the P3 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-2	HNC	*l*	HNC	PCM	2.3	2.94	4.33	1.25
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.26	2.97	4.17	1.25
0	PSE-2	HNC	*g*	HNC	PCM	2.25	2.98	4.12	1.25
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.29

Result 52: RMSE in kcal/mol calculated with the P3 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.25	3.18	3.93	1.16
0	PSE-2	HNC	*g*	HNC	PF	2.26	3.19	4.14	1.17
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.28	3.1	4.21	1.19
PCM	PSE-2	HNC	*g*	HNC	EC-RISM	2.26	3.07	3.97	1.2

Result 53: RMSE in kcal/mol calculated with the P3 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.6	3.82	4.91	3.22
DZ	PSE-2	HNC	*l*	HNC	PF	3.76	3.68	4.96	3.46
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.77	3.8	4.97	3.46
DZ	PSE-2	HNC	*l*	PSE	PCM	4.05	4.41	5.6	3.57
DZ	PSE-2	HNC	*l*	PSE	PF	4.2	4.25	5.65	3.81

Result 54: RMSE in kcal/mol calculated with the P3 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.6	3.82	4.91	3.22
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.77	3.8	4.97	3.46
DZ	PSE-2	HNC	*l*	HNC	PF	3.76	3.68	4.96	3.46
DZ	PSE-2	HNC	*l*	PSE	PCM	4.05	4.41	5.6	3.57
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	4.22	4.38	5.66	3.81

9.6 Lowest RMSE to the P4 model

Result 55: RMSE in kcal/mol calculated with the P4 model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	*g*	PSE	PF	2.51	2.96	3.98	1.96
0	PSE-2	Sim	*g*	PSE	PF	2.51	2.99	3.99	1.95
0	PSE-3	Sim	*g*	PSE	PCM	2.52	2.69	4.06	2.0
0	PSE-3	Sim	*g*	PSE	EC-RISM	2.53	2.92	4.01	1.99
0	PSE-3	Sim	*g*	HNC	PF	2.53	3.01	4.04	1.96

Result 56: RMSE in kcal/mol calculated with the P4 model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 57: RMSE in kcal/mol calculated with the P4 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	4.19	2.07
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.2	4.25	2.1
PCM	PSE-1	HNC	*g*	HNC	PF	2.69	3.24	3.87	2.24
0	PSE-1	HNC	*g*	HNC	PCM	2.71	3.2	4.29	2.12
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.71	3.19	4.46	2.05

Result 58: RMSE in kcal/mol calculated with the P4 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.71	3.19	4.46	2.05
0	PSE-1	HNC	*l*	HNC	PCM	2.73	3.21	4.51	2.06
0	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	4.19	2.07
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.2	4.25	2.1
0	PSE-1	HNC	*l*	HNC	PF	2.74	3.16	4.47	2.1

Result 59: RMSE in kcal/mol calculated with the P4 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	*g*	PSE	PF	2.5	2.97	3.98	1.94
0	PSE-3	Sim	*g*	PSE	PCM	2.5	2.69	4.05	1.98
0	PSE-2	Sim	*g*	PSE	PF	2.5	3.0	3.99	1.93
0	PSE-3	Sim	*g*	PSE	EC-RISM	2.51	2.92	4.0	1.96
0	PSE-2	Sim	*g*	PSE	EC-RISM	2.51	2.94	4.01	1.95

Result 60: RMSE in kcal/mol calculated with the P4 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.6	3.15	4.42	1.87
0	PSE-1	Sim	*g*	PSE	PF	2.53	3.06	4.18	1.87
0	PSE-1	Sim	*g*	PSE	EC-RISM	2.53	3.0	4.2	1.88
0	PSE-1	Sim	*g*	HNC	PF	2.61	3.21	4.4	1.88
0	PSE-1	Sim	*g*	PSE	PCM	2.56	3.02	4.24	1.91

Result 61: RMSE in kcal/mol calculated with the P4 model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.66	3.26	4.19	2.06
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	3.85	2.22
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.2	4.24	2.06
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.68	3.19	3.89	2.24
0	PSE-1	HNC	*g*	HNC	PCM	2.7	3.2	4.28	2.1

Result 62: RMSE in kcal/mol calculated with the P4 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.66	3.26	4.19	2.06
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.2	4.24	2.06
0	PSE-1	HNC	*g*	HNC	PCM	2.7	3.2	4.28	2.1
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	3.85	2.22
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.68	3.19	3.89	2.24

Result 63: RMSE in kcal/mol calculated with the P4 model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-3	Sim	*g*	PSE	PF	2.0	2.95	3.37	1.29
PCM	PSE-2	Sim	*g*	PSE	PF	2.0	2.94	3.41	1.26
PCM	PSE-3	Sim	*g*	PSE	PCM	2.03	2.88	3.43	1.33
PCM	PSE-3	Sim	*g*	PSE	EC-RISM	2.03	2.88	3.4	1.35
PCM	PSE-3	Sim	*g*	HNC	PF	2.03	3.01	3.36	1.32

Result 64: RMSE in kcal/mol calculated with the P4 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-3	Sim	*g*	PSE	PF	2.08	3.04	3.55	1.28
PCM	PSE-2	Sim	*g*	PSE	PF	2.08	3.03	3.57	1.26
PCM	PSE-3	Sim	*g*	PSE	EC-RISM	2.08	2.96	3.58	1.29
PCM	PSE-2	Sim	*g*	PSE	EC-RISM	2.09	2.95	3.6	1.28
PCM	PSE-3	Sim	*g*	PSE	PCM	2.1	2.96	3.62	1.3

Result 65: RMSE in kcal/mol calculated with the P4 model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	PSE	PF	2.23	2.97	4.19	1.15
0	PSE-2	HNC	*g*	HNC	PF	2.19	3.05	4.04	1.16
0	PSE-3	HNC	*g*	PSE	PF	2.21	2.6	4.27	1.2
0	PSE-3	HNC	*g*	HNC	PF	2.18	2.64	4.17	1.2
0	PSE-1	HNC	*g*	PSE	PF	2.17	3.02	3.95	1.21

Result 66: RMSE in kcal/mol calculated with the P4 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	PSE	PF	2.24	3.07	4.0	1.12
PCM	PSE-2	HNC	*g*	HNC	PF	2.2	3.14	3.81	1.13
0	PSE-2	HNC	*g*	PSE	PF	2.27	3.12	4.26	1.13
0	PSE-3	HNC	*g*	PSE	PF	2.31	3.13	4.35	1.15
D3	PSE-2	HNC	*g*	PSE	PF	2.28	3.21	3.94	1.16

Result 67: RMSE in kcal/mol calculated with the P4 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	2.06	3.06	3.52	1.26
PCM	PSE-1	HNC	*g*	HNC	PF	2.07	3.14	3.54	1.24
PCM	PSE-2	HNC	*g*	HNC	PF	2.09	3.07	3.68	1.23
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.09	3.05	3.6	1.29
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.11	2.98	3.59	1.36

Result 68: RMSE in kcal/mol calculated with the P4 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	2.15	3.13	3.71	1.29
D3	PSE-3	HNC	*g*	HNC	PF	2.18	3.3	3.88	1.24
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.19	3.06	3.78	1.35
D3	PSE-3	HNC	*g*	PSE	PF	2.2	3.24	3.96	1.23
PCM	PSE-2	HNC	*g*	HNC	PF	2.2	3.14	3.81	1.13

Result 69: RMSE in kcal/mol calculated with the P4 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	PSE	PF	2.23	2.97	4.19	1.15
0	PSE-2	HNC	*g*	HNC	PF	2.19	3.05	4.04	1.16
0	PSE-3	HNC	*g*	PSE	PF	2.21	2.6	4.27	1.2
0	PSE-3	HNC	*g*	HNC	PF	2.18	2.64	4.17	1.2
0	PSE-1	HNC	*g*	PSE	PF	2.17	3.02	3.95	1.21

Result 70: RMSE in kcal/mol calculated with the P4 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	PSE	PF	2.24	3.07	4.0	1.12
PCM	PSE-2	HNC	*g*	HNC	PF	2.2	3.14	3.81	1.13
0	PSE-2	HNC	*g*	PSE	PF	2.27	3.12	4.26	1.13
0	PSE-3	HNC	*g*	PSE	PF	2.31	3.13	4.35	1.15
D3	PSE-2	HNC	*g*	PSE	PF	2.28	3.21	3.94	1.16

Result 71: RMSE in kcal/mol calculated with the P4 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.12	3.32	4.35	2.75
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.27	3.4	4.4	2.96
DZ	PSE-2	HNC	*l*	PSE	PCM	3.27	3.36	4.63	2.88
DZ	PSE-2	HNC	*l*	HNC	PF	3.31	3.38	4.44	2.99
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	3.43	3.45	4.67	3.09

Result 72: RMSE in kcal/mol calculated with the P4 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.12	3.32	4.35	2.75
DZ	PSE-2	HNC	*l*	PSE	PCM	3.27	3.36	4.63	2.88
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.27	3.4	4.4	2.96
DZ	PSE-2	HNC	*l*	HNC	PF	3.31	3.38	4.44	2.99
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	3.43	3.45	4.67	3.09

9.7 Lowest RMSE to the P4G model

Result 73: RMSE in kcal/mol calculated with the P4G model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.65	3.18	4.39	1.96
0	PSE-1	Sim	*g*	HNC	PF	2.66	3.27	4.35	1.98
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.66	3.19	4.54	1.91
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.45	1.98
0	PSE-1	Sim	*l*	HNC	PF	2.67	3.15	4.56	1.94

Result 74: RMSE in kcal/mol calculated with the P4G model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 75: RMSE in kcal/mol calculated with the P4G model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.69	3.27	3.88	2.24
0	PSE-1	HNC	*g*	HNC	PF	2.72	3.53	4.31	2.06
D3	PSE-1	HNC	*g*	HNC	PF	2.73	3.31	3.86	2.31
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.73	3.28	3.96	2.26
D3	PSE-1	HNC	*g*	HNC	EC-RISM	2.75	3.39	3.97	2.29

Result 76: RMSE in kcal/mol calculated with the P4G model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.78	3.55	4.56	2.03
0	PSE-1	HNC	*l*	HNC	PCM	2.8	3.55	4.63	2.04
0	PSE-1	HNC	*g*	HNC	PF	2.72	3.53	4.31	2.06
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.77	3.58	4.4	2.08
0	PSE-1	HNC	*l*	HNC	PF	2.82	3.6	4.61	2.08

Result 77: RMSE in kcal/mol calculated with the P4G model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.63	3.19	4.39	1.93
0	PSE-1	Sim	*g*	HNC	PF	2.65	3.28	4.36	1.96
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.43	1.97
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.21

Result 78: RMSE in kcal/mol calculated with the P4G model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.63	3.19	4.39	1.93
0	PSE-1	Sim	*g*	HNC	PF	2.65	3.28	4.36	1.96
0	PSE-1	Sim	*g*	HNC	PCM	2.67	3.2	4.43	1.97
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05

Result 79: RMSE in kcal/mol calculated with the P4G model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.21
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05
PCM	PSE-1	HNC	*g*	HNC	PCM	2.74	3.27	3.95	2.24

Result 80: RMSE in kcal/mol calculated with the P4G model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.71	3.53	4.3	2.04
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.74	3.57	4.38	2.05
0	PSE-1	HNC	*g*	HNC	PCM	2.78	3.58	4.43	2.09
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.27	3.92	2.21
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.28	3.87	2.21

Result 81: RMSE in kcal/mol calculated with the P4G model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.39	3.67	1.32
PCM	PSE-1	Sim	*g*	HNC	PF	2.2	3.16	3.73	1.41
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	2.2	3.06	3.78	1.42
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.28
PCM	PSE-1	Sim	*g*	HNC	PCM	2.22	3.06	3.82	1.43

Result 82: RMSE in kcal/mol calculated with the P4G model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.21	3.06	3.85	1.36
PCM	PSE-1	HNC	*g*	PSE	PF	2.21	3.18	3.79	1.37
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
PCM	PSE-1	HNC	*g*	PSE	PCM	2.24	3.07	3.87	1.41
PCM	PSE-2	HNC	*g*	HNC	PF	2.25	3.18	3.93	1.16

Result 83: RMSE in kcal/mol calculated with the P4G model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.26	2.95	4.18	1.24
0	PSE-2	HNC	*g*	HNC	PCM	2.25	2.98	4.13	1.25
0	PSE-2	HNC	*l*	HNC	PCM	2.3	2.94	4.33	1.25
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.28

Result 84: RMSE in kcal/mol calculated with the P4G model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.25	3.18	3.93	1.16
0	PSE-2	HNC	*g*	HNC	PF	2.26	3.19	4.14	1.17
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.28	3.1	4.21	1.19
PCM	PSE-2	HNC	*g*	HNC	PCM	2.28	3.06	4.01	1.19

Result 85: RMSE in kcal/mol calculated with the P4G model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.18	3.39	3.67	1.32
0	PSE-1	HNC	*g*	PSE	PF	2.22	3.06	4.0	1.28
0	PSE-1	HNC	*g*	PSE	PCM	2.23	2.93	4.05	1.31
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-1	HNC	*g*	PSE	EC-RISM	2.23	2.93	4.08	1.3

Result 86: RMSE in kcal/mol calculated with the P4G model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	2.21	3.18	3.79	1.37
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	2.21	3.06	3.85	1.36
0	PSE-1	HNC	*g*	PSE	PF	2.23	3.16	4.06	1.17
PCM	PSE-1	HNC	*g*	PSE	PCM	2.24	3.07	3.87	1.41
0	PSE-1	HNC	*g*	PSE	PCM	2.25	3.09	4.11	1.23

Result 87: RMSE in kcal/mol calculated with the P4G model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.23	3.1	4.11	1.19
0	PSE-2	HNC	*g*	HNC	EC-RISM	2.26	2.95	4.18	1.24
0	PSE-2	HNC	*l*	HNC	PCM	2.3	2.94	4.33	1.25
0	PSE-2	HNC	*g*	HNC	PCM	2.25	2.98	4.13	1.25
0	PSE-2	HNC	*l*	HNC	EC-RISM	2.32	2.89	4.39	1.28

Result 88: RMSE in kcal/mol calculated with the P4G model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	g	HNC	PF	2.25	3.18	3.93	1.16
0	PSE-2	HNC	g	HNC	PF	2.26	3.19	4.14	1.17
0	PSE-1	HNC	g	PSE	PF	2.23	3.16	4.06	1.17
PCM	PSE-2	HNC	g	HNC	PCM	2.28	3.06	4.01	1.19
0	PSE-2	HNC	g	HNC	EC-RISM	2.28	3.1	4.21	1.19

Result 89: RMSE in kcal/mol calculated with the P4G model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.6	3.82	4.9	3.22
DZ	PSE-2	HNC	*l*	HNC	PF	3.73	3.67	4.94	3.42
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.77	3.8	4.97	3.45
DZ	PSE-2	HNC	*l*	PSE	PCM	3.96	4.32	5.49	3.49
DZ	PSE-2	HNC	*l*	PSE	PF	4.14	4.18	5.6	3.74

Result 90: RMSE in kcal/mol calculated with the P4G model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	3.6	3.82	4.9	3.22
DZ	PSE-2	HNC	*l*	HNC	PF	3.73	3.67	4.94	3.42
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	3.77	3.8	4.97	3.45
DZ	PSE-2	HNC	*l*	PSE	PCM	3.96	4.32	5.49	3.49
DZ	PSE-2	HNC	*l*	PSE	PF	4.14	4.18	5.6	3.74

9.8 Lowest RMSE to the P4PSE model

Result 91: RMSE in kcal/mol calculated with the P4PSE model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.2	3.95	5.46	2.3
PCM	PSE-1	Sim	*g*	HNC	PF	3.21	3.98	5.4	2.34
PCM	PSE-1	Sim	*g*	HNC	PCM	3.23	3.96	5.52	2.32
PCM	PSE-1	Sim	*l*	HNC	PCM	3.26	3.93	5.67	2.32
0	PSE-1	Sim	*l*	HNC	EC-RISM	3.28	3.97	5.95	2.15

Result 92: RMSE in kcal/mol calculated with the P4PSE model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 93: RMSE in kcal/mol calculated with the P4PSE model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	6.05	11.13	10.15	3.16
0	PSE-1	HNC	*g*	HNC	PF	6.07	11.14	10.07	3.22
0	PSE-1	HNC	*l*	HNC	PCM	6.07	11.17	10.16	3.19
0	PSE-1	HNC	*l*	HNC	PF	6.08	11.15	10.21	3.19
D3	PSE-1	HNC	*g*	HNC	PF	6.1	11.58	10.03	3.2

Result 94: RMSE in kcal/mol calculated with the P4PSE model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	6.05	11.13	10.15	3.16
D3	PSE-3	HNC	*l*	HNC	EC-RISM	6.75	19.94	8.84	3.17
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	6.17	11.16	10.4	3.17
PCM	PSE-1	HNC	*g*	HNC	PCM	6.35	11.14	10.5	3.17
D3	PSE-1	HNC	*g*	HNC	EC-RISM	6.13	11.6	10.16	3.17

Result 95: RMSE in kcal/mol calculated with the P4PSE model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.18	3.93	5.43	2.27
PCM	PSE-1	Sim	*g*	HNC	PF	3.19	3.96	5.38	2.32
PCM	PSE-1	Sim	*g*	HNC	PCM	3.21	3.94	5.49	2.31
0	PSE-1	Sim	*g*	HNC	EC-RISM	3.27	3.99	5.85	2.18
0	PSE-1	Sim	*g*	HNC	PF	3.29	4.03	5.8	2.24

Result 96: RMSE in kcal/mol calculated with the P4PSE model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	EC-RISM	3.27	3.99	5.85	2.18
0	PSE-1	Sim	*g*	HNC	PF	3.29	4.03	5.8	2.24
0	PSE-1	Sim	*g*	HNC	PCM	3.32	4.01	5.88	2.25
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.18	3.93	5.43	2.27
PCM	PSE-1	Sim	*g*	HNC	PCM	3.21	3.94	5.49	2.31

Result 97: RMSE in kcal/mol calculated with the P4PSE model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	6.07	11.14	10.06	3.2
0	PSE-1	HNC	*g*	HNC	EC-RISM	6.09	11.16	10.18	3.17
0	PSE-1	HNC	*g*	HNC	PCM	6.1	11.17	10.15	3.21
PCM	PSE-1	HNC	*g*	HNC	PF	6.13	11.15	10.25	3.17
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	6.15	11.18	10.34	3.15

Result 98: RMSE in kcal/mol calculated with the P4PSE model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	6.15	11.18	10.34	3.15
PCM	PSE-1	HNC	*g*	HNC	PCM	6.33	11.11	10.44	3.16
PCM	PSE-1	HNC	*g*	HNC	PF	6.13	11.15	10.25	3.17
0	PSE-1	HNC	*g*	HNC	EC-RISM	6.09	11.16	10.18	3.17
0	PSE-1	HNC	*g*	HNC	PF	6.07	11.14	10.06	3.2

Result 99: RMSE in kcal/mol calculated with the P4PSE model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.0	3.68	5.25	2.02
PCM	PSE-1	Sim	*g*	HNC	PF	3.0	3.74	5.19	2.04
PCM	PSE-1	Sim	*l*	HNC	EC-RISM	3.05	3.63	5.43	2.05
PCM	PSE-1	Sim	*g*	HNC	PCM	3.05	3.7	5.34	2.05
PCM	PSE-1	Sim	*l*	HNC	PCM	3.08	3.67	5.52	2.04

Result 100: RMSE in kcal/mol calculated with the P4PSE model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.17	3.77	5.63	2.06
PCM	PSE-1	Sim	*g*	HNC	PF	3.18	3.82	5.57	2.1
PCM	PSE-1	Sim	*l*	HNC	EC-RISM	3.2	3.65	5.74	2.08
PCM	PSE-1	Sim	*g*	HNC	PCM	3.22	3.79	5.71	2.07
PCM	PSE-1	Sim	*l*	HNC	PF	3.23	3.64	5.79	2.11

Result 101: RMSE in kcal/mol calculated with the P4PSE model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	3.24	3.67	6.06	2.02
0	PSE-1	Sim	*g*	HNC	EC-RISM	3.24	3.75	5.95	2.02
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.0	3.68	5.25	2.02
0	PSE-1	Sim	*g*	HNC	PF	3.24	3.81	5.88	2.04
PCM	PSE-1	Sim	*l*	HNC	PCM	3.08	3.67	5.52	2.04

Result 102: RMSE in kcal/mol calculated with the P4PSE model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	3.43	3.79	6.48	2.03
PCM	PSE-1	Sim	*l*	HNC	PCM	3.24	3.7	5.84	2.05
0	PSE-1	Sim	*g*	HNC	EC-RISM	3.46	3.97	6.44	2.05
PCM	PSE-1	Sim	*g*	HNC	EC-RISM	3.17	3.77	5.63	2.06
0	PSE-1	Sim	*l*	HNC	PF	3.46	3.77	6.51	2.07

Result 103: RMSE in kcal/mol calculated with the P4PSE model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	5.72	10.68	8.45	3.44
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	5.77	10.7	8.57	3.46
PCM	PSE-1	HNC	*g*	HNC	PCM	5.79	10.78	8.58	3.46
PCM	PSE-1	HNC	*l*	HNC	EC-RISM	5.81	10.62	8.81	3.5
0	PSE-1	HNC	*g*	HNC	PF	5.81	11.11	8.72	3.32

Result 104: RMSE in kcal/mol calculated with the P4PSE model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	5.77	10.21	8.84	3.43
0	PSE-1	HNC	*g*	HNC	PF	5.78	10.56	8.65	3.41
PCM	PSE-1	HNC	*g*	HNC	PF	5.78	10.78	8.66	3.38
0	PSE-1	HNC	*l*	HNC	PCM	5.79	10.25	8.85	3.43
0	PSE-1	HNC	*g*	HNC	PCM	5.8	10.55	8.76	3.4

Result 105: RMSE in kcal/mol calculated with the P4PSE model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	HNC	*g*	HNC	PF	6.6	16.26	7.53	3.21
0	PSE-3	HNC	*l*	HNC	EC-RISM	6.73	16.37	8.03	3.21
0	PSE-3	HNC	*g*	HNC	EC-RISM	6.67	16.41	7.7	3.22
0	PSE-3	HNC	*g*	HNC	PCM	6.66	16.41	7.6	3.23
0	PSE-3	HNC	*l*	HNC	PF	6.79	16.48	8.09	3.27

Result 106: RMSE in kcal/mol calculated with the P4PSE model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-3	HNC	*l*	HNC	EC-RISM	7.05	17.06	7.99	3.24
DZ	PSE-3	HNC	*l*	HNC	PCM	7.04	16.98	7.86	3.29
DZ	PSE-3	HNC	*l*	HNC	PF	7.11	17.17	8.08	3.29
0	PSE-3	HNC	*g*	HNC	PCM	6.61	15.76	7.61	3.29
PCM	PSE-3	HNC	*g*	HNC	EC-RISM	6.92	16.52	7.91	3.3

Result 107: RMSE in kcal/mol calculated with the P4PSE model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	6.87	15.84	9.7	3.27
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	6.9	15.8	9.85	3.28
DZ	PSE-2	HNC	*l*	HNC	PF	6.94	15.87	9.91	3.3

Result 108: RMSE in kcal/mol calculated with the P4PSE model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	6.87	15.84	9.7	3.27
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	6.9	15.8	9.85	3.28
DZ	PSE-2	HNC	*l*	HNC	PF	6.94	15.87	9.91	3.3

9.9 Lowest RMSE to the P5 model

Result 109: RMSE in kcal/mol calculated with the P5 model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	*g*	HNC	PF	2.49	2.85	3.91	1.98
0	PSE-3	Sim	*g*	HNC	PCM	2.5	2.63	3.98	2.02
0	PSE-2	Sim	*g*	HNC	PF	2.51	2.95	3.94	1.97
0	PSE-3	Sim	*g*	HNC	EC-RISM	2.51	2.83	3.94	2.01
0	PSE-1	Sim	*g*	HNC	PF	2.52	2.98	4.12	1.91

Result 110: RMSE in kcal/mol calculated with the P5 model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 111: RMSE in kcal/mol calculated with the P5 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	4.2	2.07
PCM	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	3.98	2.18
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.2	4.29	2.08
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.7	3.21	4.06	2.2
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.71	3.2	4.51	2.02

Result 112: RMSE in kcal/mol calculated with the P5 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.71	3.2	4.51	2.02
0	PSE-1	HNC	*l*	HNC	PCM	2.73	3.21	4.53	2.05
0	PSE-1	HNC	*l*	HNC	PF	2.73	3.16	4.53	2.06
0	PSE-1	HNC	*g*	HNC	PF	2.67	3.25	4.2	2.07
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.69	3.2	4.29	2.08

Result 113: RMSE in kcal/mol calculated with the P5 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-3	Sim	*g*	HNC	PF	2.48	2.85	3.9	1.96
0	PSE-3	Sim	*g*	HNC	PCM	2.49	2.63	3.98	2.0
0	PSE-3	Sim	*g*	HNC	EC-RISM	2.49	2.83	3.93	1.98
0	PSE-2	Sim	*g*	HNC	PF	2.5	2.96	3.93	1.96
0	PSE-2	Sim	*g*	HNC	EC-RISM	2.51	2.92	3.98	1.96

Result 114: RMSE in kcal/mol calculated with the P5 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	HNC	PF	2.51	2.98	4.12	1.9
0	PSE-1	Sim	*g*	HNC	EC-RISM	2.52	2.93	4.14	1.9
0	PSE-1	Sim	*g*	HNC	PCM	2.55	2.94	4.18	1.94
0	PSE-2	Sim	*g*	HNC	PF	2.5	2.96	3.93	1.96
0	PSE-2	Sim	*g*	HNC	EC-RISM	2.51	2.92	3.98	1.96

Result 115: RMSE in kcal/mol calculated with the P5 model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.66	3.25	3.95	2.16
0	PSE-1	HNC	*g*	HNC	PF	2.66	3.26	4.2	2.05
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.2	4.02	2.15
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.21	4.28	2.05
PCM	PSE-2	HNC	*g*	HNC	PCM	2.7	3.34	4.02	2.27

Result 116: RMSE in kcal/mol calculated with the P5 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	HNC	PF	2.66	3.26	4.2	2.05
0	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.21	4.28	2.05
0	PSE-1	HNC	*g*	HNC	PCM	2.7	3.2	4.3	2.09
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.67	3.2	4.02	2.15
PCM	PSE-1	HNC	*g*	HNC	PF	2.66	3.25	3.95	2.16

Result 117: RMSE in kcal/mol calculated with the P5 model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	Sim	*g*	HNC	PF	1.97	2.84	3.4	1.22
PCM	PSE-3	Sim	*g*	HNC	PF	1.98	2.83	3.43	1.25
PCM	PSE-1	Sim	*g*	HNC	PF	1.99	2.89	3.42	1.24
PCM	PSE-2	Sim	*g*	HNC	PCM	2.01	2.78	3.45	1.3
PCM	PSE-1	Sim	*g*	HNC	PCM	2.01	2.81	3.43	1.3

Result 118: RMSE in kcal/mol calculated with the P5 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	Sim	*g*	HNC	PF	1.99	2.87	3.49	1.14
PCM	PSE-3	Sim	*g*	HNC	PF	2.0	2.8	3.59	1.15
PCM	PSE-2	Sim	*g*	HNC	EC-RISM	2.01	2.81	3.5	1.22
PCM	PSE-2	Sim	*g*	HNC	PCM	2.02	2.81	3.53	1.2
PCM	PSE-3	Sim	*g*	HNC	EC-RISM	2.02	2.76	3.59	1.2

Result 119: RMSE in kcal/mol calculated with the P5 model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.19	3.06	4.02	1.16
0	PSE-2	Sim	*g*	HNC	PF	2.03	2.81	3.74	1.18
0	PSE-1	HNC	*g*	HNC	PF	2.15	3.04	3.93	1.19
PCM	PSE-1	HNC	*g*	HNC	PF	2.02	3.08	3.45	1.2
0	PSE-3	Sim	*g*	HNC	PF	2.07	2.79	3.81	1.22

Result 120: RMSE in kcal/mol calculated with the P5 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	Sim	*g*	HNC	PF	2.04	3.12	3.93	1.14
PCM	PSE-2	Sim	*g*	HNC	PF	1.99	2.87	3.49	1.14
0	PSE-3	Sim	*g*	HNC	PF	2.07	2.97	3.9	1.15
PCM	PSE-2	HNC	*g*	HNC	PF	2.19	3.16	3.77	1.15
PCM	PSE-3	Sim	*g*	HNC	PF	2.0	2.8	3.59	1.15

Result 121: RMSE in kcal/mol calculated with the P5 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.02	3.08	3.45	1.2
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.06	3.01	3.51	1.27
PCM	PSE-1	HNC	*g*	HNC	PCM	2.07	3.01	3.55	1.28
PCM	PSE-2	HNC	*g*	HNC	PF	2.08	3.13	3.59	1.24
PCM	PSE-2	HNC	*g*	HNC	EC-RISM	2.11	3.06	3.64	1.3

Result 122: RMSE in kcal/mol calculated with the P5 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	HNC	PF	2.14	3.14	3.7	1.26
PCM	PSE-1	HNC	*g*	HNC	EC-RISM	2.16	3.08	3.77	1.31
PCM	PSE-1	HNC	*g*	HNC	PCM	2.18	3.08	3.8	1.31
D3	PSE-3	HNC	*g*	HNC	PF	2.18	3.33	3.84	1.25
PCM	PSE-2	HNC	*g*	HNC	PF	2.19	3.16	3.77	1.15

Result 123: RMSE in kcal/mol calculated with the P5 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.19	3.06	4.02	1.16
0	PSE-1	HNC	*g*	HNC	PF	2.15	3.04	3.93	1.19
PCM	PSE-1	HNC	*g*	HNC	PF	2.02	3.08	3.45	1.2
0	PSE-3	HNC	*g*	HNC	PF	2.17	2.72	4.05	1.23
0	PSE-2	HNC	*l*	HNC	PCM	2.26	2.96	4.24	1.23

Result 124: RMSE in kcal/mol calculated with the P5 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	HNC	PF	2.19	3.16	3.77	1.15
0	PSE-1	HNC	*g*	HNC	PF	2.22	3.18	4.05	1.16
D3	PSE-2	HNC	*g*	HNC	PF	2.26	3.28	3.82	1.16
0	PSE-2	HNC	*g*	HNC	PF	2.26	3.18	4.14	1.16
PCM	PSE-2	HNC	*g*	HNC	EC-RISM	2.21	3.1	3.83	1.18

Result 125: RMSE in kcal/mol calculated with the P5 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	2.86	3.41	4.11	2.41
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	2.98	3.43	4.21	2.56
DZ	PSE-2	HNC	*l*	HNC	PF	3.0	3.41	4.26	2.58

Result 126: RMSE in kcal/mol calculated with the P5 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	2.86	3.41	4.11	2.41
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	2.98	3.43	4.21	2.56
DZ	PSE-2	HNC	*l*	HNC	PF	3.0	3.41	4.26	2.58

9.10 Lowest RMSE to the P9 model

Result 127: RMSE in kcal/mol calculated with the P9 model and B3LYP/6-311+G**.The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
D3	PSE-3	HNC	*l*	PSE	PCM	2.09	2.33	3.52	1.69
D3	PSE-3	HNC	*l*	HNC	PCM	2.09	2.35	3.49	1.7
D3	PSE-3	HNC	*l*	PSE	EC-RISM	2.17	2.52	3.44	1.84
D3	PSE-3	HNC	*g*	HNC	PCM	2.17	2.23	3.45	1.84
D3	PSE-3	HNC	*g*	PSE	PCM	2.17	2.21	3.48	1.83

Result 128: RMSE in kcal/mol calculated with the P2 model and B3LYP/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*l*	HNC	EC-RISM	2.64	3.19	4.65	1.83
0	PSE-1	Sim	*l*	HNC	PCM	2.69	3.21	4.74	1.86
0	PSE-1	Sim	*l*	PSE	EC-RISM	2.6	3.07	4.48	1.86
0	PSE-1	Sim	*l*	PSE	PCM	2.63	3.08	4.57	1.88
0	PSE-1	Sim	*l*	HNC	PF	2.68	3.18	4.68	1.88

Result 129: RMSE in kcal/mol calculated with the P9 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
D3	PSE-3	HNC	*l*	PSE	PCM	2.09	2.33	3.52	1.69
D3	PSE-3	HNC	*l*	HNC	PCM	2.09	2.35	3.49	1.7
D3	PSE-3	HNC	*l*	HNC	EC-RISM	2.17	2.55	3.42	1.86
D3	PSE-3	HNC	*g*	PSE	PCM	2.17	2.21	3.48	1.83
D3	PSE-3	HNC	*g*	HNC	PCM	2.17	2.23	3.45	1.84

Result 130: RMSE in kcal/mol calculated with the P9 model, B3LYP/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*l*	PSE	EC-RISM	2.33	2.88	3.92	1.68
D3	PSE-3	HNC	*l*	PSE	PCM	2.09	2.33	3.52	1.69
D3	PSE-3	HNC	*l*	HNC	PCM	2.09	2.35	3.49	1.7
0	PSE-1	HNC	*l*	HNC	EC-RISM	2.39	2.97	4.04	1.72
0	PSE-1	HNC	*g*	PSE	PF	2.29	2.94	3.66	1.72

Result 131: RMSE in kcal/mol calculated with the P9 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	Sim	*g*	PSE	PCM	2.18	2.59	3.26	1.81
PCM	PSE-2	Sim	*g*	PSE	EC-RISM	2.19	2.88	3.22	1.79
PCM	PSE-2	Sim	*g*	PSE	PF	2.2	2.93	3.22	1.78
PCM	PSE-2	HNC	*g*	PSE	PCM	2.2	2.84	3.5	1.74
PCM	PSE-2	Sim	*g*	HNC	PCM	2.21	2.63	3.32	1.83

Result 132: RMSE in kcal/mol calculated with the P9 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	Sim	*g*	PSE	EC-RISM	2.27	2.91	3.67	1.69
0	PSE-2	Sim	*g*	PSE	EC-RISM	2.24	2.84	3.57	1.7
0	PSE-1	Sim	*g*	PSE	PF	2.29	2.99	3.66	1.7
0	PSE-2	Sim	*g*	PSE	PF	2.26	2.92	3.57	1.71
0	PSE-1	HNC	*g*	PSE	PF	2.29	2.96	3.66	1.71

Result 133: RMSE in kcal/mol calculated with the P9 model, B3LYP-D3/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	PSE	PCM	2.2	2.84	3.5	1.74
PCM	PSE-2	HNC	*g*	PSE	PF	2.21	2.91	3.47	1.77
PCM	PSE-2	HNC	*g*	PSE	EC-RISM	2.21	2.83	3.45	1.79
PCM	PSE-2	HNC	*g*	HNC	PCM	2.22	2.9	3.48	1.76
PCM	PSE-2	HNC	*g*	HNC	EC-RISM	2.23	2.9	3.43	1.81

Result 134: RMSE in kcal/mol calculated with the P9 model and B3LYP-D3/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-1	HNC	*g*	PSE	EC-RISM	2.28	2.89	3.67	1.71
0	PSE-1	HNC	*g*	PSE	PF	2.29	2.96	3.66	1.71
0	PSE-2	HNC	*g*	PSE	EC-RISM	2.29	2.8	3.68	1.74
0	PSE-2	HNC	*g*	PSE	PF	2.3	2.87	3.69	1.74
PCM	PSE-2	HNC	*g*	PSE	PCM	2.2	2.84	3.5	1.74

Result 135: RMSE in kcal/mol calculated with the P9 model and MP2/6-311+G**. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-3	Sim	*g*	PSE	PF	1.89	2.95	3.08	1.21
PCM	PSE-1	Sim	*g*	PSE	PF	1.89	2.99	3.13	1.18
PCM	PSE-2	Sim	*g*	PSE	PF	1.89	2.94	3.13	1.19
PCM	PSE-1	HNC	*g*	PSE	PF	1.89	2.95	3.1	1.18
PCM	PSE-3	Sim	*g*	HNC	PF	1.9	3.01	3.07	1.23

Result 136: RMSE in kcal/mol calculated with the P9 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-3	HNC	*l*	PSE	PCM	1.98	2.84	3.34	1.25
PCM	PSE-3	HNC	*g*	PSE	PF	1.98	2.97	3.34	1.2
D3	PSE-3	HNC	*g*	PSE	PF	1.99	3.11	3.37	1.21
PCM	PSE-3	HNC	*l*	PSE	PCM	1.99	2.84	3.33	1.28
PCM	PSE-2	HNC	*g*	PSE	PF	1.99	2.98	3.3	1.1

Result 137: RMSE in kcal/mol calculated with the P9 model and MP2/6-311+G**. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	Sim	*g*	PSE	PF	1.93	2.97	3.39	1.12
0	PSE-2	HNC	*g*	HNC	PF	2.01	2.96	3.51	1.14
0	PSE-2	HNC	*g*	PSE	PF	1.99	2.87	3.5	1.14
0	PSE-2	Sim	*g*	HNC	PF	1.96	3.06	3.43	1.14
0	PSE-3	Sim	*g*	PSE	PF	1.95	2.97	3.41	1.16

Result 138: RMSE in kcal/mol calculated with the P9 model and MP2/aug-cc-pVDZ. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	PSE	PF	1.99	2.98	3.3	1.1
0	PSE-2	HNC	*g*	PSE	PF	2.04	3.02	3.6	1.12
D3	PSE-2	HNC	*g*	PSE	PF	2.07	3.1	3.39	1.13
PCM	PSE-2	HNC	*g*	HNC	PF	2.03	3.06	3.34	1.13
0	PSE-3	HNC	*g*	PSE	PF	2.04	3.02	3.58	1.13

Result 139: RMSE in kcal/mol calculated with the P9 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-1	HNC	*g*	PSE	PF	1.89	2.95	3.1	1.18
PCM	PSE-2	HNC	*g*	PSE	PF	1.91	2.9	3.16	1.22
PCM	PSE-2	HNC	*g*	HNC	PF	1.92	2.98	3.14	1.22
PCM	PSE-1	HNC	*g*	PSE	EC-RISM	1.92	2.88	3.11	1.26
PCM	PSE-1	HNC	*g*	PSE	PCM	1.92	2.88	3.12	1.26

Result 140: RMSE in kcal/mol calculated with the P9 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-3	HNC	*l*	PSE	PCM	1.98	2.84	3.34	1.25
PCM	PSE-3	HNC	*g*	PSE	PF	1.98	2.97	3.34	1.2
PCM	PSE-2	HNC	*g*	PSE	PF	1.99	2.98	3.3	1.1
D3	PSE-3	HNC	*g*	PSE	PF	1.99	3.11	3.37	1.21
PCM	PSE-3	HNC	*l*	PSE	PCM	1.99	2.84	3.33	1.28

Result 141: RMSE in kcal/mol calculated with the P9 model, MP2/6-311+G** and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
0	PSE-2	HNC	*g*	HNC	PF	2.01	2.96	3.51	1.14
0	PSE-2	HNC	*g*	PSE	PF	1.99	2.87	3.5	1.14
0	PSE-1	HNC	*g*	PSE	PF	1.99	2.93	3.49	1.16
PCM	PSE-1	HNC	*g*	PSE	PF	1.89	2.95	3.1	1.18
0	PSE-3	HNC	*g*	PSE	PF	1.95	2.51	3.51	1.19

Result 142: RMSE in kcal/mol calculated with the P9 model, MP2/aug-cc-pVDZ and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
PCM	PSE-2	HNC	*g*	PSE	PF	1.99	2.98	3.3	1.1
0	PSE-2	HNC	*g*	PSE	PF	2.04	3.02	3.6	1.12
D3	PSE-2	HNC	*g*	PSE	PF	2.07	3.1	3.39	1.13
0	PSE-3	HNC	*g*	PSE	PF	2.04	3.02	3.58	1.13
PCM	PSE-2	HNC	*g*	HNC	PF	2.03	3.06	3.34	1.13

Result 143: RMSE in kcal/mol calculated with the P9 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of all molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	HNC	PCM	2.28	2.82	3.63	1.77
DZ	PSE-2	HNC	*l*	PSE	PCM	2.29	2.77	3.71	1.75
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	2.3	2.81	3.6	1.82
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	2.31	2.76	3.68	1.81
DZ	PSE-2	HNC	*l*	HNC	PF	2.34	2.76	3.61	1.89

Result 144: RMSE in kcal/mol calculated with the P9 model, B3LYP/6-31+G* and the HNC solvent model. The data is sorted by RMSE of neutral molecules.

Geom	closure	χ	ref.	μ	ensemble	all	cations	anions	neutrals
DZ	PSE-2	HNC	*l*	PSE	PCM	2.29	2.77	3.71	1.75
DZ	PSE-2	HNC	*l*	HNC	PCM	2.28	2.82	3.63	1.77
DZ	PSE-2	HNC	*l*	PSE	EC-RISM	2.31	2.76	3.68	1.81
DZ	PSE-2	HNC	*l*	HNC	EC-RISM	2.3	2.81	3.6	1.82
DZ	PSE-2	HNC	*l*	PSE	PF	2.35	2.71	3.68	1.88

10 References

[1] Ikeguchi, M.; Doi, J. *The Journal of Chemical Physics* 1995, *103*, 5011.

[2] Beglov, D.; Roux, B. *The Journal of Chemical Physics* 1995, *103*, 360.

[3] Kovalenko, A.; Hirata, F. *Chemical Physics Letters* 1998, *290*, 237-244.

[4] Beglov, D.; Roux, B. *The Journal of Chemical Physics* 1996, *104*, 8678.

[5] Kloss, T.; Heil, J.; Kast, S. *The Journal of Physical Chemistry B* 2008, *112*, 4337-4343.

[6] Widom, B. *The Journal of Chemical Physics* 1963, *39*, 2808.

[7] Henderson, J. *Molecular Physics* 1983, *48*, 389-400.

[8 Lee, L.; Shing, K. *The Journal of Chemical Physics* 1989, *91*, 477.

[9] Barboy, B.; Tennne, R. *Chemical Physics* 1979, *38*, 369-387.

[10] Kelly, C.; Cramer, C.; Truhlar, D. *J. Chem. Theory Comput.* 2005, *1*, 1133-1152.

[11] Cramer, C.; Truhlar, D. *Accounts of Chemical Research* 2008, *41*, 760-768.

[12] Liu, J.; Kelly, C.; Goren, A.; Marenich, A.; Cramer, C.; Truhlar, D.; Zhan, C. *J. Chem. Theory Comput.* 2010, *6*, 1109-1117.

[13] Morita, T.; Hiroike, K. *Prog. Theor. Phys.* 1961, *25*, 537-578.

[14] Kovalenko, A.; Hirata, F. *The Journal of Chemical Physics* 2000, *112*, 10403.

[15] Kast, S.; Kloss, T. *The Journal of Chemical Physics* 2008, *129*, 236101.

[16] Kast, S.; Tomazic, D. *The Journal of Chemical Physics* 2012, *137*, 171102.

[17] Ben-Naim, A. *Solvation thermodynamics*; Plenum Press: New York, 1987.

[18] Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database – version 2012*; University of Minnesota, Minneapolis, 2012

[19] Kast, S.; Heil, J.; Güssregen, S.; Schmidt, K. *Journal of Computer-Aided Molecular Design* 2010, *24*, 355-355.

[20] Ho, J.; Klamt, A.; Coote, M. *J. Phys. Chem. A* 2010, *114*, 13442-13444.

[21] Ribeiro, R.; Marenich, A.; Cramer, C.; Truhlar, D. *The Journal of Physical Chemistry B* 2011, *115*, 14556-14562.

[22] Tissandier, M.; Cowen, K.; Feng, W.; Gundlach, E.; Cohen, M.; Earhart, A.; Coe, J.; Tuttle, T. *J. Phys. Chem. A* 1998, *102*, 7787-7794.

[23] Liptak, M.; Shields, G. *J. Am. Chem. Soc.* 2001, *123*, 7314-7319.

[24] Heil, J.; Tomazic, D.; Egbers, S.; Kast, S. *Journal of Molecular Modeling* 2014, *20*.

[25] Coe, J. *Chemical Physics Letters* 1994, *229*, 161-168.

[26] Coe, J. *International Reviews in Physical Chemistry* 2001, *20*, 33-58.

[27] MacKerell, A.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W.; Roux, B.;

Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *The Journal of Physical Chemistry B* 1998, *102*, 3586-3616.

[28] See the implementation of the CHARMM pair potential in the LAMMPS MD software http://lammps.sandia.gov/doc/pair_charmm.html

[29] Mulero, A. *Theory and simulation of hard-sphere fluids and related systems*; Springer: Berlin, 2008.

[30] Nicolas, J.; Gubbins, K.; Streett, W.; Tildesley, D. *Molecular Physics* 1979, *37*, 1429-1454.

[31] Johnson, J.; Zollweg, J.; Gubbins, K. *Molecular Physics* 1993, *78*, 591-618.

[32] Weeks, J. *The Journal of Chemical Physics* 1971, *54*, 5237.

[33] Duh, D.; Haymet, A. *The Journal of Chemical Physics* 1995, *103*, 2625.

[34] Choudhury, N.; Ghosh, S. *The Journal of Chemical Physics* 2002, *116*, 8517.

[35] Mederos, L.; Navascués, G.; Tarazona, P. *Physical Review E* 1994, *49*, 2161-2166.

[36] Berendsen, H.; Grigera, J.; Straatsma, T. *The Journal of Physical Chemistry* 1987, *91*, 6269-6271.

[37] Maw, S.; Sato, H.; Ten-no, S.; Hirata, F. *Chemical Physics Letters* 1997, *276*, 20-25.

[38] Sato, H.; Hirata, F. *The Journal of Chemical Physics* 1999, *111*, 8545.

[39] Lue, L.; Blankschtein, D. *The Journal of Physical Chemistry* 1992, *96*, 8582-8594.

[40] Beutler, T.; Mark, A.; van Schaik, R.; Gerber, P.; van Gunsteren, W. *Chemical Physics Letters* 1994, *222*, 529-539.

[41] Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossva ry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, .; Ye, X.; Wang, J.; Hsieh, M. J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Solomon-Ferrer, R.; Sagui, C.; abin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. Amber12 ; University of California: San Francisco, CA, 2012

[42] Mennucci, B.; Cammi, R. *Front Matter, in Continuum solvation models in chemical physics*; John Wiley & Sons: Chichester, England, 2007.

[43] Tomasi, J.; Mennucci, B.; Cammi, R. *Chemical Reviews* 2005, *105*, 2999-3094.

[44] Amovilli, C.; Barone, V.; Cammi, R.; Cancès, E.; Cossi, M.; Mennucci, B.; Pomelli, C.; Tomasi, J. *Advances in Quantum Chemistry* 1998, 227-261.

[45] Qiu, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* 1997, *101*, 3005-3014.

[46] Sigalov, G.; Fenley, A.; Onufriev, A. *The Journal of Chemical Physics* 2006, *124*, 124902.

[47] Kelly, C.; Cramer, C.; Truhlar, D. *The Journal of Physical Chemistry B* 2006, *110*, 16066-16081.

[48] Marenich, A.; Olson, R.; Kelly, C.; Cramer, C.; Truhlar, D. *J. Chem. Theory Comput.* 2007, *3*, 2011-2033.

[49] Marenich, A.; Cramer, C.; Truhlar, D. *J. Chem. Theory Comput.* 2009, *5*, 2447-2464.

[50] Cancès, E.; Mennucci, B.; Tomasi, J. *The Journal of Chemical Physics* 1997, *107*, 3032.

[51] Cancès, E.; Mennucci, B. *Journal of Mathematical Chemistry* 1998, *23*, 309-326.

[52] Mennucci, B.; Cancès, E.; Tomasi, J. *The Journal of Physical Chemistry B* 1997, *101*, 10506-10517.

[53] Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* 1993, 799.

[54] Klamt, A. *The Journal of Physical Chemistry* 1995, *99*, 2224-2235.

[55] Andreussi, O.; Dabo, I.; Marzari, N. *The Journal of Chemical Physics* 2012, *136*, 064102.

[56] Dupont, C.; Andreussi, O.; Marzari, N. *The Journal of Chemical Physics* 2013, *139*, 214110.

[57] Sundararaman, R.; Goddard, W. *The Journal of Chemical Physics* 2015, *142*, 064107.

[58] Hansen, J.; McDonald, I. *Theory of simple liquids*; Elsevier Academic Press: London, 2006.

[59] Evans, R. *Advances in Physics* 1979, *28*, 143-200.

[60] Hirata, F. *Molecular theory of solvation*; Kluwer Academic Publishers: Dordrecht, 2003.

[61] Blum, L. *The Journal of Chemical Physics* 1972, *56*, 303.

[62] Richardi, J.; Millot, C.; Fries, P. *The Journal of Chemical Physics* 1999, *110*, 1138.

[63] Kovalenko, A.; Hirata, F. *The Journal of Chemical Physics* 2000, *112*, 10391.

[64] Morita, T.; Hiroike, K. *Progress of Theoretical. Physics.* 1960, *23*, 385-387.

[65] Andersen, H. *The Journal of Chemical Physics* 1972, *57*, 1918.

[66] Chandler, D. *The Journal of Chemical Physics* 1972, *57*, 1930.

[67] Beglov, D.; Roux, B. *The Journal of Physical Chemistry B* 1997, *101*, 7821-7826.

[68] Ratkova, E.; Palmer, D.; Fedorov, M. *Chemical Reviews* 2015, *115*, 6312-6356.

[69] Hirata, F.; Rossky, P. *Chemical Physics Letters* 1981, *83*, 329-334.

[70] Perkyns, J.; Pettitt, B. *The Journal of Chemical Physics* 1992, *97*, 7656.

[71] Perkyns, J.; Montgomery Pettitt, B. *Chemical Physics Letters* 1992, *190*, 626-630.

[72] Chuev, G.; Vyalov, I.; Georgi, N. *Chemical Physics Letters* 2013, *561-562*, 175-178.

[73] Kovalenko, A.; Hirata, F. *The Journal of Physical Chemistry B* 1999, *103*, 7942-7957.

[74] Kovalenko, A.; Hirata, F. *The Journal of Chemical Physics* 1999, *110*, 10095.

[75] Verlet, L. *Molecular Physics* 1980, *41*, 183-190.

[76] Duh, D.; Haymet, A. *The Journal of Chemical Physics* 1992, *97*, 7716.

[77] Duh, D.; Henderson, D. *The Journal of Chemical Physics* 1996, *104*, 6742.

[78] Kast, S. *Physical Review E* 2003, *67*.

[79] Lado, F. *Physics Letters A* 1982, *89*, 196-198.

[80] Kolafa, J.; Labík, S.; Malijevský, A. *Molecular Physics* 2002, *100*, 2629-2640.

[81] Kinoshita, M.; Imai, T.; Kovalenko, A.; Hirata, F. *Chemical Physics Letters* 2001, *348*, 337-342.

[82] Hiroike, K. *Journal of the Physical Society of Japan* 1960, *15*, 771-778.

[83] Iyetomi, H. *Progress of Theoretical Physics* 1984, *71*, 427-437.

[84] Kovalenko, A.; Hirata, F.; Kinoshita, M. *The Journal of Chemical Physics* 2000, *113*, 9830.

[85] Vyalov, I.; Chuev, G.; Georgi, N. *The Journal of Chemical Physics* 2014, *141*, 074505.

[86] Rossky, P.; Dale, W. *The Journal of Chemical Physics* 1980, *73*, 2457.

[87] Ciccotti, G.; Ferrario, M.; Hynes, J.; Kapral, R. *Chemical Physics* 1989, *129*, 241-251.

[88] Kambayashi, S.; Chihara, J. *Physical Review E* 1994, *50*, 1317-1324.

[89] Llano-Restrepo, M.; Chapman, W. *The Journal of Chemical Physics* 1992, *97*, 2046.

[90] Chiles, R.; Rossky, P. *J. Am. Chem. Soc.* 1984, *106*, 6867-6868.

[91] Singer, S.; Chandler, D. *Molecular Physics* 1985, *55*, 621-625.

[92] Kjellander, R.; Sarman, S. *The Journal of Chemical Physics* 1989, *90*, 2768.

[93] Vompe, A.; Martynov, G. *The Journal of Chemical Physics* 1994, *100*, 5249.

[94] Lomba, E.; Lee, L. *International Journal of Thermophysics* 1996, *17*, 663-672.

[95] Imai, T.; Kinoshita, M.; Hirata, F. *The Journal of Chemical Physics* 2000, *112*, 9469.

[96] Harano, Y.; Imai, T.; Kovalenko, A.; Kinoshita, M.; Hirata, F. *The Journal of Chemical Physics* 2001, *114*, 9506.

[97] Kirkwood, J.; Buff, F. *The Journal of Chemical Physics* 1949, *17*, 338.

[98] Kežić, B.; Perera, A. *The Journal of Chemical Physics* 2011, *135*, 234104.

[99] Du, Q.; Beglov, D.; Roux, B. *The Journal of Physical Chemistry B* 2000, *104*, 796-805.

[100] Kast, S. *Phys. Chem. Chem. Phys.* 2001, *3*, 5087-5092.

[101] Kast, S.; Friedemann Schmidt, K.; Schilling, B. *Chemical Physics Letters* 2003, *367*, 398-404.

[102] Puibasset, J.; Belloni, L. *The Journal of Chemical Physics* 2012, *136*, 154503.

[103] Zhao, S.; Liu, H.; Ramirez, R.; Borgis, D. *The Journal of Chemical Physics* 2013, *139*, 034503.

[104] Chuev, G.; Vyalov, I.; Georgi, N. *J. Comput. Chem.* 2014, *35*, 1010-1023.

[105] Plimpton, S. *Journal of Computational Physics* 1995, *117*, 1-19.

[106] Nosé, S. *The Journal of Chemical Physics* 1984, *81*, 511.

[107] Hoover, W. *Phys. Rev. A* 1985, *31*, 1695-1697.

[108] Hirschfelder, J.; Curtiss, C.; Bird, R. *Molecular theory of gases and liquids*; Wiley: New York, 1954.

[109] Ryckaert, J.; Ciccotti, G.; Berendsen, H. *Journal of Computational Physics* 1977, *23*, 327-341.

[110] Talman, J. *Journal of Computational Physics* 1978, *29*, 35-48.

[111] Kovalenko, A.; Ten-no, S.; Hirata, F. *J. Comput. Chem.* 1999, *20*, 928.

[112] Lee, L. *The Journal of Chemical Physics* 1992, *97*, 8606.

[113] Kast, S. personal communication

[114] Carnahan, N. *The Journal of Chemical Physics* 1969, *51*, 635.

[115] Allen, M.; Tildesley, D. *Computer simulation of liquids*; Clarendon Press: Oxford [England], 1987.

[116] Dierckx, P. *Journal of Computational and Applied Mathematics* 1975, *1*, 165-184.

[117] Dierckx, P. *SIAM Journal on Numerical Analysis* 1982, *19*, 1286-1304.

[118] Guthrie, J. *The Journal of Physical Chemistry B* 2009, *113*, 4501-4507.

[119] Nicholls, A.; Mobley, D.; Guthrie, J.; Chodera, J.; Bayly, C.; Cooper, M.; Pande, V. *J. Med. Chem.* 2008, *51*, 769-779.

[120] Skillman, A.; Geballe, M.; Nicholls, A. *Journal of Computer-Aided Molecular Design* 2010, *24*, 257-258.

[121] Mobley, D.; Wymer, K.; Lim, N.; Guthrie, J. *Journal of Computer-Aided Molecular Design* 2014, *28*, 135-150.

[122] Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R.; Sherman, W. *J. Chem. Theory Comput.* 2012, *8*, 2553-2558.

[123] Fennell, C.; Kehoe, C.; Dill, K. *Proceedings of the National Academy of Sciences* 2011, *108*, 3234-3239.

[124] Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. *Accounts of Chemical Research* 2009, *42*, 489-492.

[125] Chambers, C.; Hawkins, G.; Cramer, C.; Truhlar, D. *The Journal of Physical Chemistry* 1996, *100*, 16385-16398.

[126] Li, J.; Hawkins, G.; Cramer, C.; Truhlar, D. *Chemical Physics Letters* 1998, *288*, 293-298.

[127] Li, J.; Zhu, T.; Hawkins, G.; Winget, P.; Liotard, D.; Cramer, C.; Truhlar, D. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* 1999, *103*, 9-63.

[128] Frach, R.; Kast, S. *J. Phys. Chem. A* 2014, *118*, 11620-11628.

[129] Palmer, D.; Frolov, A.; Ratkova, E.; Fedorov, M. *Journal of Physics: Condensed Matter* 2010, *22*, 492101.

[130] Palmer, D.; Frolov, A.; Ratkova, E.; Fedorov, M. *Mol. Pharmaceutics* 2011, *8*, 1423-1429.

[131] Ng, K. *The Journal of Chemical Physics* 1974, *61*, 2680.

[132] Sergiievskyi, V.; Jeanmairet, G.; Levesque, M.; Borgis, D. *J. Phys. Chem. Lett.* 2014, *5*, 1935-1942.

[133] Sergiievskyi, V.; Jeanmairet, G.; Levesque, M.; Borgis, D. *The Journal of Chemical Physics* 2015, *143*, 184116.

[134] Misin, M.; Fedorov, M.; Palmer, D. *The Journal of Physical Chemistry B* 2016.

[135] Imai,. *Condensed Matter Physics* 2007, *10*, 343.

[136] Becke, A. *Phys. Rev. A* 1988, *38*, 3098-3100.

[137] Lee, C.; Yang, W.; Parr, R. *Phys. Rev. B* 1988, *37*, 785-789.

[138] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *The Journal of Chemical Physics* 2010, *132*, 154104.

[139] Møller, C.; Plesset, M. *Phys. Rev.* 1934, *46*, 618-622.

[140] Ditchfield, R. *The Journal of Chemical Physics* 1971, *54*, 724.

[141] Dunning, T. *The Journal of Chemical Physics* 1989, *90*, 1007.

[142] Kendall, R.; Dunning, T.; Harrison, R. *The Journal of Chemical Physics* 1992, *96*, 6796.

[143] Blaney, J.; Dixon, J. *Reviews in Computational Chemistry* 299-335.

[144] Havel, T.; Kuntz, I.; Crippen, G. *Bulletin of Mathematical Biology* 1983, *45*, 665-720.

[145] Spellmeyer, D.; Wong, A.; Bower, M.; Blaney, J. *Journal of Molecular Graphics and Modelling* 1997, *15*, 18-36.

[146] Engel, J.; Richters, A.; Getlik, M.; Tomassi, S.; Keul, M.; Termathe, M.; Lategahn, J.; Becker, C.; Mayer-Wrangowski, S.; Grütter, C.; Uhlenbrock, N.; Krüll, J.; Schaumann, N.; Eppmann, S.; Kibies, P.; Hoffgaard, F.; Heil, J.; Menninger, S.; Ortiz-Cuaran, S.; Heuckmann, J.; Tinnefeld, V.; Zahedi, R.; Sos, M.; Schultz-Fademrecht, C.; Thomas, R.; Kast, S.; Rauh, D. *J. Med. Chem.* 2015, *58*, 6844-6863.

[147] Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. *J. Comput. Chem.* 2004, *25*, 1157-1174.

[148] Wang, J.; Wang, W.; Kollman, P.; Case, D. *Journal of Molecular Graphics and Modelling* 2006, *25*, 247-260.

[149] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. *J. Comput. Chem.* 2005, *26*, 1701-1718.

[150] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision D.02*; Gaussian, Inc., Wallingford CT, 2009.

[151] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gom-

perts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.02*; Gaussian, Inc., Wallingford CT, 2004.

[152] Floriano, W.; Nascimento, M. *Brazilian Journal of Physics* 2004, *34*, 38-41.

[153] Chirlian, L.; Francl, M. *J. Comput. Chem.* 1987, *8*, 894-905.

[154] Grimme, S. *J. Comput. Chem.* 2006, *27*, 1787-1799.

[155] Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* 2011, *32*, 1456-1465.

[156] Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* 2008, *4*, 435-447.

[157] Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. *The Journal of Chemical Physics* 1995, *103*, 8577.

[158] Miyamoto, S.; Kollman, P. *J. Comput. Chem.* 1992, *13*, 952-962.

[159] Klicić, J.; Friesner, R.; Liu, S.; Guida, W. *J. Phys. Chem. A* 2002, *106*, 1327-1335.

[160] Schleyer, P. *Encyclopedia of computational chemistry*; New York: Chichester, 1998.

[161] Havel, T. F. *Distance Geometry: Theory, Algorithms, and Chemical Applications, in Encyclopedia of computational chemistry*; New York: Chichester, 2002.