Physicochemical property prediction for small molecules using integral equation-based solvation models

Dissertation zur Erlangung des
akademischen Grades eines Doktors der
Naturwissenschaften (Dr. rer. nat.)

Die Dissertation wurde im Zeitraum vom 01.11.15 bis zum 09.02.21 angefertigt und der Fakultät Chemie und Chemische Biologie der Technischen Universität Dortmund vorgelegt

von
Nicolas Tielker.

Dortmund, 2021

Erstgutachter:	Prof. Dr. Stefan M. Kast
Zweitgutachter:	Prof. Dr. Paul Czodrowski

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

1: N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, S. M. Kast, *J. Comput.-Aided Mol. Des*. 30, 1035 (2016)

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

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

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

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

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

Danksagungen

Ich möchte mich zuerst herzlich bei Herrn Prof. Dr. Stefan M. Kast bedanken, durch dessen unermüdliche Unterstützung in fachlichen Dingen und die Schaffung einer angenehmen Arbeitskreisatmosphäre, ich immer für meine wissenschaftliche Arbeit und die Lehre motiviert war. Auch für das Interesse an der Situation der Doktoranden nach ihrer Promotion, insbesondere im Hinblick auf das vergangene Coronajahr möchte ich mich herzlich bedanken. Bei Herrn Prof. Dr. Paul Czodrowski bedanke ich mich für die Übernahme des Zweitgutachtens und bei Herrn Dr. Stefan Güssregen für die vielen fruchtvollen Kooperationen im Bereich der „SAMPL challenges" und den anregenden Gesprächen auf so mancher Konferenz.

Darüber hinaus möchte ich mich bei allen Mitgliedern des Arbeitskreises für die gute Zusammenarbeit und Arbeitsatmosphäre danken, wo einem bei so gut wie jedem Problem von mindestens einer Person geholfen wurde, und auch die sportliche Betätigung nie zu kurz kam. Besonders danken möchte ich dabei Dr. Daniel Tomazic, Dr. Roland Frach, Dr. Leonhard Henkes, Dr. Jochen Heil und Dr. Florian Mrugalla, die, gerade in der Übergangsphase vom Master in die Promotion, mit Rat und Tat geholfen haben. Dr. Patrick Kibies, der nicht nur bei fachlichen Themen oder Fragen zur Administration und zum Coding, sondern auch in vielen anderen Bereichen mit nützen und unnützen Wissen glänzen konnte. Martin Urban für die Unterstützung insbesondere bei MD-Simulationen und der Lehre. Lukas Eberlein und Tim Pongratz für die angenehme Atmosphäre im Büro zu Beginn der Promotion, die sich auch mit der im Laufe der Zeit wechselnden Besetzung erhalten hat. und insbesondere nochmal bei Lukas Eberlein, Julia Jasper, Tim Pongratz, und Patrick Kibies für die sehr angenehme Gesellschaft bei diversen Konferenzen, Workshops, und anderen Dienstreisen. Bei Frau Anneliese Ahlke bedanke ich mich besonders für die scheinbar mühelose Bewältigung aller Verwaltungsangelegenheiten, durch die das wissenschaftliche Arbeiten sehr viel leichter von der Hand ging.

Schließlich möchte ich mich noch ganz besonders bei meinen Eltern, Birgit Popiolek und Uwe Tielker bedanken. Das Wissen, dass es jemanden gibt auf den man sich in allen Lebenslagen verlassen kann, hat mich im Laufe meines Studiums von vielen kleinen und großen Sorgen befreit. Auch Martina Tielker und Klaus Zipfel habe einen großen Beitrag zu dieser Sicherheit geleistet.

Abstract

This thesis is concerned with the accurate prediction of physicochemical properties of small, pharmaceutically relevant compounds. As they are closely related to the pharmacokinetic profile, knowing these properties prior to expending money and resources on the molecules' synthesis is of considerable interest to pharmaceutical companies [7,8,9]. In addition, these properties can be measured experimentally and used as benchmarks to compare the accuracy of predictions by different theoretical methods. To predict these condensed phase properties such as hydration free energies, acid dissociation constants (pK_a), and distribution and partition coefficients (log D and log P, respectively) it is necessary to accurately describe the solute, the solute-solvent interactions, and the solvent-response to the solute's presence. When this is achieved, the Gibbs energies of the molecules in solution can be used to directly calculate these macroscopic properties.

The embedded cluster reference interaction site model (EC-RISM) makes it possible to combine a quantum chemical (QC) description of the solute with an accurate solvent response via the three-dimensional reference interaction site model (3D RISM) [10,11,12,13]. This is ideal for calculating physicochemical properties of small molecules, because EC-RISM yields both the electronic energy of the solvent-polarized wave function, as well as the excess chemical potential of the molecule in solution, the sum of which can be defined as the Gibbs energy of the molecule in solution. The combination of the very accurate QC description of the solute with the quick calculation of the equilibrium solvent structure and excess chemical potential makes it possible to treat a large number of compounds with good accuracy.

The calculation of Gibbs energies with EC-RISM requires a number of preparatory steps. On the solvent side, the solvent response function, also called solvent susceptibility, must be pre-computed for use in 3D RISM. For the organic solvents necessary to calculate partition coefficients between immiscible phases these susceptibilities did not exist when this thesis was started. The calculation of the solvent susceptibilities requires suitable Lennard-Jones parameters and the equilibrium structure of the solvent atoms. Additionally, the partially water-miscible solvent octanol might not be accurately described by a pure octanol phase and thus there was a necessity for generating solvent susceptibilities of water-octanol mixtures. On the solute side, proper conformational sampling is necessary because the intramolecular energy of the solute strongly depends on the molecule's geometry. Furthermore, the solute electrostatics must be accurately extracted from the wave function so they can be used in the 3D RISM calculations.

To summarize the following work, in this thesis the development of solvent susceptibilities for the non-aqueous solvents cyclohexane and *n*-octanol is reported, as well as the challenges and implications of including water saturation for organic solvents. The solvent susceptibilities are then used to train partial molar volume (PMV) corrections to correct for the error inherent in the calculation of the 3D RISM excess chemical potential using reference data from the Minnesota solvation database (MNSOL) [14,15,16,17]. Additionally, a method to calculate accurate pK_a values is presented and the formal equivalence of a microstate transition and a partition function approach is briefly summarized. The performance of the models is benchmarked by participation in the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges. The first application was in the SAMPL5 challenge, where cyclohexane-water distribution coefficients log $D_{7.4}$ had to be calculated [1]. This task pushed the limits of what the theoretical community was capable of and even the best participants had unsatisfactory results. For this reason, in the subsequent challenges the task was split into determining aqueous pK_a values during the SAMPL6 challenge [2] and octanol-water partition coefficients log P of a subset of these compounds for SAMPL6 part II [5]. Over the course of these challenges a number of key improvements were made to the EC-RISM model, often directly as a result of inconsistencies or performance issues during one of the SAMPL challenges. These will be reported in detail in the respective chapters. Finally, an extension of the partial molar volume correction to extreme conditions such as high pressure is reported.

Ongoing EC-RISM developments based on this work involve the application of EC-RISM to other solvent conditions, solvent mixtures, or aqueous electrolytes through the development of PMV corrections, force field reparametrization, and combination with machine learning methods.

ZUSAMMENFASSUNG

Die vorliegende Arbeit behandelt die genaue Vorhersage physikochemischer Parameter von kleinen, pharmazeutisch relevanten Molekülen. Da diese Eigenschaften eng mit dem pharmakokinetischen Profil des Stoffes zusammenhängen ist es von großem Interesse für Pharmaunternehmen diese bereits vor der kosten- und zeitaufwändigen Synthese der Substanz zu kennen. Zudem können diese Eigenschaften experimentell bestimmt und als Maßstab für den Vergleich der Genauigkeit verschiedener theoretischer Methoden verwendet werden. Um diese Eigenschaften wie Freie Hydratationsenthalpien, Säurekonstanten (pK_a) und Verteilungskoeffizienten (log D und log P), die stark vom Lösungsmittel abhängen, genau vorhersagen zu können ist es notwendig sowohl das Solvat, die Solvat-Lösungsmittel-Interaktionen und den Effekt des Solvats auf das Lösungsmittel zu berücksichtigen. Dann ist es möglich die Gibbs-Energien der Moleküle in Lösung zu verwenden, um diese makroskopischen Eigenschaften zu berechnen.

Das *embedded cluster reference interaction site model* (EC-RISM) ermöglicht es eine quantenchemische Beschreibung des Solvats mit einer genauen Beschreibung des Effekts des Solvats auf das Solvens in Form des dreidimensionalen *reference interaction site models* (3D RISM) zu kombinieren. Dies ist vorteilhaft für die Berechnung von physikochemischen Parametern, denn EC-RISM liefert sowohl die elektronische Energie der lösungsmittelpolarisierten Wellenfunktion als auch das chemische Exzesspotential des Moleküls in Lösung. Die Summe dieser beiden Größen kann als Gibbs-Energie des Moleküls in Lösung definiert werden. Durch die Kombination einer sehr genauen quantenchemischen Beschreibung des gelösten Teilchens mit schnell zu berechnenden Gleichgewichts-Lösungsmittelstrukturen und chemischen Exzesspotentialen ermöglicht es eine große Anzahl von Verbindungen mit hoher Genauigkeit zu behandeln.

Die Berechnung der Gibbs-Energien mit EC-RISM benötigt einige vorbereitende Schritte. Auf der Lösungsmittelseite muss die sogenannte Lösungsmittelsuszeptibilität, welche die Antwort des Lösungsmittels auf eine Störung codiert, vorberechnet werden damit sie in 3D RISM-Rechnungen verwendet werden kann. Für die organischen Lösungsmittel die notwendig sind, um Verteilungskoeffizienten zwischen zwei nicht mischbaren Phasen zu berechnen existierten zu Beginn dieser Arbeit noch keine dieser Funktionen. Die Berechnung dieser Funktionen erfordert geeignete Lennard-Jones-Parameter für die Interaktionszentren und eine Gleichgewichtsstruktur des Lösungsmittelmoleküls. Zudem ist zu bedenken, dass die Octa-

nolphase durchaus Wasser beinhalten kann, und deshalb die Darstellung als reine Octanolphase das Experiment nicht hinreichend beschreiben kann. Aus diesem Grund ist es auch nötig die Lösungsmittelsuszeptibilitäten von Octanol-Wasser-Mischungen zu generieren. Auf Seite des gelösten Teilchens müssen die energetisch günstigsten Konformationen durch konformationelles sampling gefunden werden, da die intramolekulare Energie stark von der Molekülgeometrie abhängt. Außerdem muss die Elektrostatik des gelösten Teilchens mit hinreichender Genauigkeit aus der Wellenfunktion extrahiert werden, damit sie in den RISM-Rechnungen verwendet werden kann.

In dieser Arbeit wird die Entwicklung von Lösungsmittelsuszeptibilitäten für die nichtwäßrigen Lösungsmittel Cyclohexan und *n*-Octanol beschrieben. Auch die Herausforderungen und Folgen der Modellierung von mit Wasser gesättigten organischen Lösungsmitteln wird diskutiert. Diese Lösungsmittelsuszeptibilitäten werden dann genutzt um Korrekturen für den 3D RISM inhärenten Fehler bei der Berechnung des chemischen Exzesspotentials an der *Minnesota solvation database* (MNSOL) zu trainieren, welche das partielle molare Volumen der Solvate als Parameter nutzen. Zudem wird eine Methode zur Berechnung genauer Säurekonstanten vorgestellt und für die Vorhersage von pK_a-abhängigen Verteilungskoeffizienten genutzt. Die Qualität der Modelle wird durch Teilnahme an den *Statistical Assessment of Modeling of Proteins and Ligand challenges* (SAMPL *challenges*) überprüft und dabei mit anderen modernen Methoden zur Bestimmung dieser physikochemischen Eigenschaften verglichen. Zunächst wurden die Modelle während der SAMPL5 *challenge* eingesetzt, um Cyclohexan-Wasser-Distributionskoeffizienten log $D_{7.4}$ zu berechnen. Diese Aufgabe war so komplex, dass selbst die besten Modelle, die an der *challenge* teilnahmen, keine zufriedenstellenden Ergebnisse erzielen konnten. Aus diesem Grund wurde diese Aufgabe für die SAMPL6 *challenge* in zwei getrennten Teilen behandelt. Zunächst mussten nur die Säurekonstanten (pKa) bestimmt werden, bevor in einem zweiten Teil, „SAMPL6 part II“, die Octanol-Wasser-Verteilungskoeffizienten log *P* bestimmt werden mussten. Im Verlauf dieser *challenges* wurden eine Reihe von wichtigen Verbesserungen am EC-RISM-Modell gemacht, häufig als Antwort auf Inkonsistenzen oder schlechten Ergebnissen während einer der SAMPL *challenges*. Schließlich wird eine Erweiterung der Korrektur für das chemische Exzesspotential für Extrembedingungen wie hohen hydrostatischen Druck vorgestellt.

Die Weiterentwicklung von EC-RISM mit den Methoden, die in dieser Arbeit entwickelt wurden, beinhaltet unter anderem die Anwendung von EC-RISM auf andere thermodynamische Bedingungen, Lösungsmittelmischungen und wässrige Elektrolyte. Dies wird durch die

Entwicklung neuer PMV-Korrekturen, Kraftfeldreparametrisierungen und Kombination mit Methoden des maschinellen Lernens erreicht.

TABLE OF CONTENTS

1 Introduction ... 12

1.1 Motivation ... 12

1.2 Aims and objectives ... 18

2 Theoretical Background ... 20

2.1 The embedded cluster reference interaction site model EC-RISM ... 20

2.1.1 Classical density functional theory ... 20

2.1.2 1D RISM and the generation of solvent susceptibilities ... 24

2.1.3 The 3D RISM solvation model ... 25

2.1.4 EC-RISM ... 28

2.2 Calculation of physicochemical properties ... 29

2.2.1 Empirical correction of the excess chemical potential ... 29

2.2.2 Calculation of acidity constants ... 31

2.2.3 Partition and distribution coefficients ... 33

2.3 Molecular dynamics simulations ... 35

2.3.1 Introduction ... 35

2.3.2 Liquid structure from MD simulations ... 36

2.3.3 Thermodynamic integration ... 38

3 Generation of solvent susceptibilities ... 39

3.1 Introduction ... 39

3.1.1 Cyclohexane ... 39

3.1.2 Octanol and water-octanol mixtures ... 39

3.2 Computational details ... 40

3.2.1 Cyclohexane MD simulation ... 40

3.2.2 Solvent susceptibilities and 1D RISM calculations ... 41

3.3 Results and discussion ... 42
3.3.1 Cyclohexane ... 42
3.3.2 Wet *n*-octanol ... 43
4 SAMPL5: Calculation of cyclohexane-water distribution coefficients log $D_{7.4}$ ... 45
4.1 Introduction ... 45
4.2 Computational details ... 46
4.3 Results ... 48
4.3.1 Solvation free energies ... 48
4.3.2 pK_a prediction ... 51
4.3.3 Prediction of distribution coefficients ... 52
5 SAMPL6.1: prediction of acidity constants in aqueous solution ... 58
5.1 Introduction ... 58
5.2 Computational details ... 59
5.3 Results ... 60
5.3.1 Solvation free energies ... 60
5.3.1.1 MP2-based models ... 61
5.3.1.2 DFT-based models ... 64
5.3.2 pK_a model training ... 67
5.3.2.1 MP2-based models ... 68
5.3.2.2 DFT-based models ... 76
5.3.3 pK_a model application ... 78
5.3.4 Prediction of pH-dependent tautomer ratios ... 85
6 SAMPL6.2: prediction of partition coefficients between water and octanol ... 89
6.1 Introduction ... 89
6.2 Computational details ... 90
6.3 Results ... 90
6.3.1 Solvation free energies ... 90

6.3.2 MNSOL partition coefficients ... 92

6.3.3 SAMPL6.2 partition coefficients ... 93

6.4 SAMPL5 revisited ... 98

6.4.1 Introduction ... 98

6.4.2 Computational Details ... 99

6.4.3 Results ... 100

7 Partial Molar Volume correction for molecules under extreme conditions ... 110

7.1 Introduction ... 110

7.2 Computational details ... 112

7.3 Results ... 113

8 Summary and conclusion ... 117

9 References ... 120

1 Introduction

1.1 Motivation

The increasing cost of developing new drugs has led to increased attention to all parts of the drug development pipeline [18,19]. Especially in the preclinical stage of development not just the biological efficacy, but also the pharmacokinetic profile of new lead molecules is of great importance. The pharmacokinetic profile is usually characterized with the acronym ADME(T) or ADME-Tox which stands for absorption, distribution, metabolism, excretion and toxicity [20,21]. These properties must be properly balanced when a compound is developed as a drug and each of those areas can lead to failure in the drug development process. For example, the ideal drug formulation enables oral or topical application, but sufficient bioavailability is more difficult to achieve chemically than for intravenously administered drugs [22]. The distribution of a compound usually takes place through the blood stream, and biological obstacles such as the blood-brain barrier can make some targets even more difficult to address. The metabolism is closely related to the toxicity and the excretion of the molecule, as some of the metabolic products may have adverse reactions in the human body, but even just too quick or too slow metabolic clearance can make a drug too difficult to dose for widespread use. These properties are not characterized by just a single or a simple combination of physical properties but are the result of the highly complex interplay between a compound and the biological systems in the human body.

Some physicochemical properties can be estimated more easily than others. For example, even though the actual biology involved can be more complex, a simple property such as the partition coefficient between water and octanol, the log P_{ow}, can be used as a measure of a molecule's oral bioavailability, which involves the absorption and distribution of a compound in the human body [7,8,9]. Generally speaking, a drug candidate must not be too hydrophilic, because it would not be able to cross through biological membranes that consist of lipid bilayers, but it also must not be too lipophilic, as that may lead to accumulation in those bilayers [23]. Furthermore, high lipophilicity is implicated in several processes that lead to failure during drug development such as toxic side effects, binding to plasma proteins and general target promiscuity [24,25,26,27]. This example shows how important it is for drug researchers to accurately determine a drug candidate's hydrophilicity and to know how to modulate it.

This simple model can be refined by using a more accurate measure, the log $D_{7.4}$, that further takes the molecule's protonation state at the physiological pH of 7.4 into consideration [28]. Compounds that contain an ionizable group can have a membrane permeability that differs significantly from what the partition coefficient would indicate, because the protonated or deprotonated species are much less likely to enter the membrane. This effect cannot be captured by the log P because it, by definition, only measures the neutral species' partitioning. The log $D_{7.4}$ is more difficult to predict because it requires knowledge of the acidity constants of the molecule in question, but it is also a more accurate measure when profiling pharmaceutically relevant compounds for their *in vivo* hydrophilicity. For most molecules that contain ionizable groups with a pK_a that is far from the physiological pH, i.e. very weak acids and bases that are predominantly found in their neutral forms at pH 7.4, the partition coefficient is effectively equal to the distribution coefficient. For those with an ionizable group outside of that range the effects can be extremely strong, possibly even shifting a molecule that would be considered very lipophilic from its neutral state partition coefficient log P to very hydrophilic when looking at the log $D_{7.4}$.

The protonation state of molecules at physiological pH is of a more general interest as well. A protein interacts with ligands in different ways, but often large contributions to the Gibbs free energy of binding are made by hydrogen bonding and electrostatic interactions such as salt bridges [29]. The protonation or deprotonation of a ligand can change the hydrogen bonding pattern accessible to the protein and introduce charged groups leading to increased or lowered affinity, depending on the system under investigation.

To gain access to all parts of the pharmacokinetic profile a multitude of methods have been developed over the course of the late 20th and early 21st century to correlate physicochemical properties to experimentally measurable physical properties, the three-dimensional molecular structure, or molecular descriptors derived from the structure, such as the number and type of heteroatoms, the number of hydrogen bond acceptors and donors, or the polar surface area [30,31]. This approach called quantitative structure-activity relationship (QSAR) when investigating efficacy and quantitative structure-property relationship (QSPR) when investigating e.g. pharmacokinetic parameters. In recent years, these approaches have been bolstered using deep learning and similar artificial intelligence (AI) methods that take advantage of the progress made in high performance computing and data science [32,33]. The routine use of computational methods to predict the pharmacokinetic profile has only recently become possible through the improved computational power of modern computer clusters and the increased availability of large, highly-curated training sets, even though the correlation of e.g. partition coefficients and membrane diffusion rates have been known for a long time [34].

To deal with the enormous size of chemical space, i.e. the diversity of potentially synthesizable molecules, modern drug research relies heavily on the existence of compound libraries [35]. Due to their large size, however, in a typical molecular library the pharmacokinetic properties of most compounds may not be known experimentally, but only by computational methods. Furthermore, during the lead optimization phase it can be beneficial to know the change in the pharmacokinetic properties of a proposed optimization before spending time and resources on synthesizing the compound and measuring the experimental values. In all those cases the use of computational methods, be they empirical or physical in nature can be a great help for the modern medicinal chemist. For most compounds, the quality of predictions made using empirical methods are sufficient and more expensive methods that are unfeasible to use for libraries consisting of millions of compounds are not required. However, one drawback of these empirical methods is that they only work reliably in that part of chemical space that they were parametrized in. New chemical moieties that are of great interest to many pharmaceutical companies may require retraining or additional empirical corrections [36].

Partition and distribution coefficients as well as acidity constants have in common that they either are only relevant in solution (partition and distribution coefficients) or at least have a strong dependence on the solvent (acidity constants). Thus, for accurate predictions of such properties the effect of the solvent on the solute must be modeled accurately as well. Various solvation models have been developed to model solvation in computational approaches: continuum models such as the polarizable continuum model (PCM) and generalized Born solvation models (GB/SA) or even explicit solvation using atomistic solvent molecules in molecular dynamics (MD) or *ab initio* molecular dynamics (aiMD) simulations [37,38,39]. The continuum models are comparatively fast, because no explicit solute-solvent atom interactions must be calculated and so they are very well suited to handle large numbers of molecules. Explicit solvation on the other hand requires larger computational resources than their continuum counterparts, but if properly used the accuracy can be significantly greater due to the explicit modeling of interactions at an atomistic level such as hydrogen bonding and solvent structure disruptions [40,41,42]

Compared to continuum solvent models and explicit solvation the 3D reference interaction site model (3D RISM) makes it possible to determine the equilibrium solvent structure of a given solvent around the solute with significantly less computational effort than using atomistic simulations would require [11,12]. This alleviates some of the issues that are inherent to many implicit solvent models, namely that the structural information of the solvent is lost and with it the thermodynamic contribution to the solvation process.

Combining this solvent model with a quantum-chemical description of the solvent makes it possible to accurately determine both the intramolecular energies and the solute-solvent interactions of a given compound [10]. This embedded-cluster reference interaction site model (EC-RISM) had already been applied on a number of tasks related to physicochemical property prediction, such as the prediction of tautomer ratios, relative acidity constants or the Gibbs energies of hydration of the molecules in the Minnesota Solvation Database (MNSOL) [17,43,44]. However, at the beginning of this work, EC-RISM had not been applied on a large scale to calculate more complex physicochemical properties such as absolute acidity constants or partition coefficients.

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) series of challenges [45] originated in 2008 from a collaboration of the scientific software company OpenEye with scientists from Stanford University. This first challenge, later termed SAMPL0, consisted of the blind prediction of 17 small molecule aqueous solvation free energies [46]. An important decision that was made when designing this challenge, and that is now one of the cornerstones of the SAMPL challenges, was the attempt to set it up as a “blind” prediction challenge. This means that the data that is to be predicted has not been previously published and thus cannot have been used as reference data e.g. to develop the predictive models used during the challenge. For the challenges from SAMPL0 to

SAMPL4 the goal of creating a blind challenge was not yet fully achieved because the property to predict, hydration free energies, are difficult to measure [46,47,48,49,50]. Instead, obscure sources such as vapor pressures or boiling point data were used to calculate experimental solvation free energies of compounds that were not part of readily available solvation free energy databases [50]. Eventually in 2015, the SAMPL5 challenge was issued, and the participants were given the task to predict cyclohexane-water distribution coefficients of a set of 53 organic molecules [51]. This new target property has the advantage of being much more experimentally accessible, as it can be determined using conventional LC-MS/MS (liquid chromatography with tandem mass spectrometry) methods that are commonly used in biochemistry labs [52,53]. This means that the challenges could be conducted on compounds specifically designed for the SAMPL challenges and avoiding any kind of already published data even if it is sufficiently obscure. From a computational point of view on the other hand this new target property adds to the complexity of the challenge. To calculate the partition coefficient between water and cyclohexane in this challenge requires on the one hand the development of an entirely new solvent model for cyclohexane, as this solvent had never been used with EC-RISM before. Additionally, it would be necessary to apply the existing water models in a new area because the calculation of distribution coefficients requires knowledge of the compounds' acid dissociation constant for ionizable compounds. The big jump in complexity of this challenge compared to the preceding ones led to an adjustment in the SAMPL6 challenge. Here, the computation of the distribution coefficients was split into separate challenges for the acidity constants pK_a [54] and the partition coefficients between water and octanol [55]. This pattern was also used in the recently finished SAMPL7 challenge [56].

Blind challenges such as the SAMPL series of challenges are designed to facilitate the comparison of wildly different methods of predicting a given molecular property. By generating but not publishing the experimental results in advance the predicted values are as unbiased as possible, because the methods could not have been trained using the same compounds. Furthermore, the publication of all results regardless of their quality is very valuable. In many disciplines scientific papers too often focus only on positive results due to funding pressures or perhaps even the prestige associated with success [57]. Failures and negative results are sidelined and in extreme cases never become known outside of the group that produced them. This leads to redundant research when other groups repeat the same failed experiments and turns science into a chase for the best numbers instead of knowledge. Only by publishing the good and the bad results, as is done in the SAMPL series of challenges, can the reasons for e.g. their relative performance be analyzed and, ideally, further insight into the natural phenomena under investigation be gained.

1.2 Aims and objectives

Prior to the development of this work, EC-RISM had been applied to a variety of problems, including the calculation of relative Gibbs energies of tautomers in aqueous solution [43]. However, before the development and testing of empirical corrections to the 3D RISM excess chemical potential by D. Tomazic it was not possible to calculate properties such as solvation free energies [44]. The reason for that is a systematic error in the 3D RISM excess chemical potential that is highly correlated with the partial molar volume (PMV) of the solute for reasons that will be explained in more detail in Chapter 2.2.1. This work is intended to show the step-by-step extension of this approach to be applicable to non-aqueous solvents and the improvements developed for water solvent models.

Early during this work, the SAMPL5 challenge on predicting cyclohexane-water distribution coefficients for small, drug-like molecules was issued. These blind challenges were considered to be a great opportunity to apply the PMV correction and test its performance in a real world setting against other approaches from various areas of computational modeling. The SAMPL5 challenge additionally required the development of a cyclohexane model for use with EC-RISM and parametrization of a PMV correction for it. For this solvent, the solvent susceptibility function was generated by directly inverting the total correlation function extracted from a molecular dynamics (MD) simulation, which had not been done before. In addition, a model to calculate acidity constants from the Gibbs energies of the neutral and the charged species had to be developed. This first task and the results of the SAMPL5 challenge are detailed in chapters 3 and 4, however at that point in time only a subset of the molecules could be investigated due to time constraints. Effectively splitting the task of the SAMPL5 challenge into two separate parts, the next challenge, SAMPL6.1, was concerned only with aqueous acidity constants. This made it possible to test a variety of quantum chemical levels of theory, basis sets, PMV correction approaches, pK_a models, and approaches to model the solute electrostatics for the water model to further improve upon those developed by D. Tomazic. The results of these extensive tests are summarized in chapter 5. Following that, the neutral state octanol-water partition coefficients for a subset of the compounds investigated had to be predicted during the SAMPL6.2 challenge. For this, the solvent model for octanol had to be generated, but, unlike in the case of cyclohexane, a significant, experimentally measurable amount of water can be dissolved in the octanol phase. For this reason, both a neat and a water-saturated octanol model were generated (chapter 3) and applied to the challenge compounds to predict the partition coefficients (chapter 6). This marks the first time the PMV correction approach for EC-RISM was applied to a solvent mixture. Finally, the improvements made over the course of the SAMPL challenges were reapplied to the compounds of the SAMPL5 challenge to determine the effect of the improvements (chapter 6.4). This time the entire set of compounds measured in the original challenge except for one could be investigated.

In addition to these properties under normal conditions a workflow was developed that makes it possible to use the partial molar volume correction for solvents even under extreme conditions (chapter 7) to enable the investigation of absolute molecular properties under high hydrostatic pressure using EC-RISM. In this work the focus lies on high pressure but repurposing of the workflow to e.g. high temperatures is also possible. Due to a lack of available experimental data this workflow uses the change in the Gibbs energy of solvation at 1 bar compared to higher pressures as calculated using thermodynamic integration (TI) as the reference during the parametrization.

2 Theoretical Background

2.1 The embedded cluster reference interaction site model EC-RISM

2.1.1 Classical density functional theory

Density functional theory (DFT) is well-known for electronic systems where according to the Hohenberg-Kohn theorem of quantum DFT the energy of a quantum system is uniquely determined by the density distribution of the electrons [58]. Similarly, however, even for classical systems the Helmholtz energy A of a fluid is a unique functional of the single particle density $\rho^{(1)}$. The following derivation is a summary of the full derivation of the so-called Ornstein-Zernike equation for molecular fluids that can be pieced together from the literature [59,60,61,62].

Given an arbitrary external potential V_{ext} the grand potential of the grand canonical ensemble is given by

$$\Omega_V\left[\rho\right] = \int d\mathbf{r}\rho(\mathbf{r})V_{\text{ext}}(\mathbf{r}) + A\left[\rho\right] - \mu\int d\mathbf{r}\rho(\mathbf{r}) \tag{1}$$

where A is the Helmholtz energy, $\mathbf{r}$ the three-dimensional coordinates, and μ the excess chemical potential. Square brackets denote functionals while round brackets denote functions. It can be shown that the density that minimizes the functional in eq. (1) is the equilibrium density ρ_0, i.e.

$$\left.\frac{\delta\Omega_V\left[\rho\right]}{\delta\rho(\mathbf{r})}\right|_{\rho_0}=0 \tag{2}$$

and thus

$$\Omega_V\left[\rho_0\right]=\Omega. \tag{3}$$

Knowing this the total Helmholtz energy A can be defined as the sum of the intrinsic Helmholtz energy and the contribution of the external potential

$$A=\int d\mathbf{r}\rho_0(\mathbf{r})V_{\text{ext}}(\mathbf{r})+A\left[\rho_0\right] \tag{4}$$

and the excess chemical potential arises from the fundamental thermodynamic relation of the Helmholtz energy as

$$\mu=V_{\text{ext}}(\mathbf{r})+\frac{\delta A\left[\rho_0\right]}{\delta\rho(\mathbf{r})}=V_{\text{ext}}(\mathbf{r})+\mu_{\text{in}}[\rho_0;\mathbf{r}] \tag{5}$$

with μ_{in} as the intrinsic chemical potential that arises only from the particle interactions.

Furthermore, it is also possible to define the total Helmholtz energy as a sum of the contribution arising from an ideal, non-interacting system and a functional for the interaction part $\varphi\left[\rho\right]$ as

$$A\left[\rho\right]\equiv A_{\text{id}}\left[\rho\right]-\varphi\left[\rho\right] \tag{6}$$

From the definition of the intrinsic chemical potential as the functional derivative of the Helmholtz free energy with regards to the density in eq. (5) it follows that

$$\beta\frac{\delta A\left[\rho\right]}{\delta\rho(\mathbf{r})}=\beta\mu_{\text{in}}[\rho;\mathbf{r}]=\ln\left(\Lambda^3\rho(\mathbf{r})\right)-c[\rho;\mathbf{r}] \tag{7}$$

with the thermal wavelength $\Lambda=\sqrt{2\pi\beta\hbar^2/m}$, the thermodynamic beta $\beta=1/(k_BT)$, $\hbar$ as the reduced Planck constant, m as the mass, and where

$$c[\rho;\mathbf{r}]\equiv\beta\frac{\delta\varphi[\rho]}{\delta\rho[\mathbf{r}]} \tag{8}$$

is the direct correlation function that includes all contributions caused by particle interactions. Higher order direct correlation functions can be generated by differentiating the direct correlation function again, most importantly the second order direct correlation function

$$c^{(2)}[\rho;\mathbf{r}_1,\mathbf{r}_2]\equiv\frac{\delta c[\rho;\mathbf{r}_1]}{\delta\rho[\mathbf{r}_2]}=\frac{\delta\varphi[\rho]}{\delta\rho[\mathbf{r}_1]\delta\rho[\mathbf{r}_2]} \tag{9}$$

"is also called the Ornstein-Zernike direct correlation function of the non-uniform fluid" [60]. One additional important relation that can be derived from first principles in this way is the grand-canonical density-density correlation function

$$\begin{aligned} H^{(2)}(\mathbf{r},\mathbf{r}') &= \left\langle \left[\rho(\mathbf{r}) - \left\langle \rho(\mathbf{r}) \right\rangle\right]\left[\rho(\mathbf{r}') - \left\langle \rho(\mathbf{r}') \right\rangle\right] \right\rangle \\ &= \rho^{(1)}(\mathbf{r})\rho^{(1)}(\mathbf{r}')h^{(2)}(\mathbf{r},\mathbf{r}') + \rho^{(1)}(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}') \end{aligned} \tag{10}$$

where the particle density of order *n* is

$$\rho^{(n)}\left(\mathbf{r}^{n}\right) = \frac{1}{\Xi}\sum_{N=n}^{\infty}\frac{1}{(N-n)!}\int d\mathbf{r}^{(N-n)}\exp\left(-\beta V_{N}\right)\left(\prod_{i=1}^{N} z\exp\left[-\beta V_{\text{ext}}(\mathbf{r}_{i})\right]\right) \tag{11}$$

with V_N as the interatomic potential energy, z as a local activity $z = \exp\left(-\beta\mu\right)\Lambda^{-3}$ and the grand canonical partition function

$$\Xi = \sum_{N=0}^{\infty}\frac{1}{N!}\int d\mathbf{r}^{N}\exp\left(-\beta V_{N}\right)\left(\prod_{i=1}^{N} z\exp\left[-\beta V_{\text{ext}}(\mathbf{r}_{i})\right]\right). \tag{12}$$

$h^{(2)}(\mathbf{r},\mathbf{r}')$ is the total correlation function that is related to the liquid structure by way of the radial distribution function as

$$h^{(2)}(\mathbf{r},\mathbf{r}') = g(\mathbf{r},\mathbf{r}') - 1\,. \tag{13}$$

To finally derive the Ornstein-Zernike equation it is necessary to take the second functional derivative of the grand canonical potential with respect to the intrinsic chemical potential which yields the expression

$$\frac{\delta^{2}\Omega_{V}}{\delta\mu_{\text{in}}(\mathbf{r})\delta\mu_{\text{in}}(\mathbf{r}')} = \beta^{-1}\frac{\delta\rho_{0}(\mathbf{r})}{\delta\mu_{\text{in}}(\mathbf{r}')} = H^{(2)}(\mathbf{r},\mathbf{r}') \tag{14}$$

and the inverse of this functional derivative is

$$\left(H^{(2)}(\mathbf{r},\mathbf{r}')\right)^{-1} = \beta\frac{\delta\mu_{\text{in}}(\mathbf{r}')}{\delta\rho_{0}(\mathbf{r})}. \tag{15}$$

Comparison of this result to the definition of the second order direct correlation function defined in eq. (9) as the functional derivative of the direct correlation function and the reformulating eq. (7) as

$$c[\rho;\mathbf{r}] = \ln\left(\Lambda^{3}\rho(\mathbf{r})\right) - \beta\mu_{\text{in}}[\rho;\mathbf{r}]. \tag{16}$$

yields

$$c^{(2)}(\mathbf{r},\mathbf{r}') \equiv c^{(2)}[\rho_{0};\mathbf{r},\mathbf{r}'] = \frac{\delta(\mathbf{r}-\mathbf{r}')}{\delta\rho_{0}(\mathbf{r})} - \beta\frac{\delta\mu_{\text{in}}(\mathbf{r}')}{\delta\rho_{0}(\mathbf{r})} = \frac{\delta(\mathbf{r}-\mathbf{r}')}{\delta\rho_{0}(\mathbf{r})} - \left(H^{(2)}(\mathbf{r},\mathbf{r}')\right)^{-1}. \tag{17}$$

By combining eq. (10) and eq. (17) with the definition of the functional inverse

$$\int d\mathbf{r}''H^{(2)}(\mathbf{r},\mathbf{r}'')\left(H^{(2)}(\mathbf{r}'',\mathbf{r}')\right)^{-1} = \delta(\mathbf{r}-\mathbf{r}'). \tag{18}$$

and integrating over $\mathbf{r}''$ it is possible to derive the Ornstein-Zernike equation

$$h^{(2)}(\mathbf{r},\mathbf{r}') = c^{(2)}(\mathbf{r},\mathbf{r}') +_{0} \int d\mathbf{r}''h^{(2)}(\mathbf{r}'',\mathbf{r}')\rho^{(1)}(\mathbf{r}'')c^{(2)}(\mathbf{r},\mathbf{r}''). \tag{19}$$

that connects the direct correlation function $c^{(2)}(\mathbf{r},\mathbf{r'})$ and the total correlation function $h^{(2)}(\mathbf{r},\mathbf{r'})$ [59,60].

This is an equation with two unknowns, the direct and the total correlation function, so a second relation between these is needed to solve it, the so-called closure relation. It is possible to derive an approximate closure relation for the Ornstein-Zernike equation from the definition of the grand potential in eq. (1): By expressing the intrinsic Helmholtz energy in powers of $\Delta\rho$ coupled to a reference system of density ρ_0, chemical potential μ_0 and the same temperature, and setting the resulting coupling-parameter dependent direct correlation function equal to the direct correlation function of the reference system the intrinsic free energy is approximately

$$A \approx A_0 + \mu_0^{\text{in}} \int d\mathbf{r}\Delta\rho(\mathbf{r}) - \frac{1}{2}\beta^{-1}\int d\mathbf{r}d\mathbf{r'}\Delta\rho(\mathbf{r})c_0(\mathbf{r},\mathbf{r'})\Delta\rho(\mathbf{r'}) \tag{20}$$

where the index “0” denotes the reference system. It is possible to show that the density which minimizes eq. (20) is

$$\rho(\mathbf{r}) = \rho_0 \exp\left(-\beta V_{\text{ext}}(\mathbf{r}) + \int d\mathbf{r'}\Delta\rho(\mathbf{r'})c_0(|\mathbf{r}-\mathbf{r'}|)\right). \tag{21}$$

and together with the Ornstein-Zernike equation this yields

$$g(\mathbf{r}) = \exp\left(h(\mathbf{r}) - c(\mathbf{r}) - \beta u(\mathbf{r})\right). \tag{22}$$

which represents the hypernetted-chain closure (HNC) for the Ornstein-Zernike equation [59].

To extend this theory to molecular systems the molecules must be described through both the coordinates of the molecular center $\mathbf{R}_i$ and the molecular orientation expressed by its Euler angles $\mathbf{\Omega}_i \equiv (\theta_i, \varphi_i, \chi_i)$. For ease of notation each molecule i is then defined through the unique combination $i \equiv (\mathbf{R}_i, \mathbf{\Omega}_i)$ and the molecular Ornstein-Zernike equation can be written by generalizing the atomic Ornstein-Zernike equation as

$$h(1,2) = c(1,2) + \frac{\rho_0}{\Omega}\int c(1,3)h(3,2)d3, \tag{23}$$

where the numbers are the 6-dimensional representation of the molecules [59,63]. The differential $d3$, too, indicates integration over the coordinates and Euler angles. It is possible to eliminate the total correlation function from the integral by inserting eq. (23) into itself, leading to the infinite sum expansion

$$h(1,2) = c(1,2) + \frac{\rho_0}{\Omega}\int c(1,3)c(3,2)d3 + \frac{\rho_0^2}{\Omega}\int c(1,3)c(3,4)c(4,2)d4 + \ldots,. \tag{24}$$

which is useful for further derivations that are described in more detail in the next chapter. Other approaches, such as diagrammatic representations of the OZ equation are not discussed here but can be found in the literature [59,64].

2.1.2 1D RISM and the generation of solvent susceptibilities

Since the resulting 6-dimensional Ornstein-Zernike equation is difficult to use in a productive setting some approximate theories have been developed. The reference interaction site model is based on describing the molecule as a set of rigidly connected hard spheres in which the spheres may overlap [65]. In the most basic form, the spheres correspond to the molecule's atoms, but this is not strictly required by the theory and united atom approaches that model entire chemical groups as one sphere are possible. The derivation of the RISM equation is based on an approximation of the direct correlation function as the sum of all site-site direct correlation functions

$$c(\mathbf{r}) \approx \sum_{\alpha}\sum_{\gamma} c_{\alpha\gamma}(\mathbf{r}_{\alpha\gamma}) \tag{25}$$

where α and γ are the solute and solvent sites, respectively. If the direct correlation function can be decomposed into site-site correlation function the same is true for the total correlation function. This can be proven by substituting the Fourier transform of eq. (25) into the Fourier transform of the molecular OZ equation, i.e. eq. (24), and averaging over the orientations of the two molecules [61,62,65]. This yields the RISM integral equation

$$\rho h_{\alpha\gamma}(r) = \sum_{\gamma'}\sum_{\alpha'}\int\int \omega_{\alpha\alpha'}(|\mathbf{r}_1-\mathbf{r}'|)\times c_{\alpha'\gamma'}(|\mathbf{r}'-\mathbf{r}''|)\chi_{\gamma\gamma'}(|\mathbf{r}''-\mathbf{r}_2|)d\mathbf{r}'d\mathbf{r}'' \tag{26}$$

where $\chi_{\gamma\gamma'}$ is the (pure) solvent susceptibility function that is defined as

$$\chi_{\gamma\gamma'}(r) = \rho\omega_{\gamma\gamma'}(r) + \rho h_{\gamma\gamma'}(r)\rho \tag{27}$$

$\omega_{\alpha\alpha'}$ is the intramolecular correlation function that has a similar physical interpretation as the intermolecular correlation function but between sites in the same molecule. It can thus be considered a measure for the equilibrium molecular structure as described through average site-site distances, as compared to the equilibrium solvent structure described by the intermolecular correlation function. It is usually approximated as

$$\omega_{\alpha\alpha'}(r) = \frac{\delta(r-r_{\alpha\alpha'})}{4\pi r_{\alpha\alpha'}{}^{2}} \tag{28}$$

which implies a completely rigid molecule, because the small molecular vibrations of e.g. water are known to have little effect on the results but other approximations that include molecular flexibility are possible [66]. The correlation between sites in different molecules is then governed by both the intra- and intermolecular correlations between the different sites.

The solvent susceptibility can be calculated self-consistently by calculating a solvent molecule solvated by itself as solvent with a suitable closure relation. It is also possible to extract the inter- and, if necessary, intramolecular distribution function from MD simulations and calculate the solvent susceptibility matrix elements using these functions [67,68]. This is achieved by using a modified HNC closure of the form

$$h(r)+1=\exp[h(r)-c(r)-\beta u(r)+B^{\mathrm{MD}}(r)] \tag{29}$$

where B^{MD} is constructed in such a way that it constrains the total correlation functions from RISM calculations to the simulated total correlation functions calculated from the radial distribution functions for short distances. To still retain the correct HNC long-range behavior B^{MD} contains a cubic polynomial switching function that varies between 1 and 0 over a predefined range so that

$$B^{\mathrm{MD}}(r)=f(r)\left[\ln\left(g^{\mathrm{MD}}(r)\right)-h(r)+c(r)+\beta u(r)\right] \tag{30}$$

with $f(r)$ as the switching function and $g^{\mathrm{MD}}(r)$ as the radial distribution function extracted from MD simulation. The solvent susceptibilities generated this way can then be used for 3D RISM calculations [1,69].

2.1.3 The 3D RISM solvation model

3D RISM is an extension of 1D RISM theory where only the molecular orientations of the solvent atoms are averaged out of the six-dimensional Ornstein-Zernike equation, while those of the solute are explicitly considered [11,62,70,71]. The partial averaging over the molecular orientations yields the 3D RISM equation

$$\rho h_{\gamma}(\mathbf{r})=\sum_{\gamma'}\int\int c_{\gamma'}(|\mathbf{r}-\mathbf{r}'|)\chi_{\gamma\gamma'}(|\mathbf{r}'|)\mathrm{d}\mathbf{r}' \tag{31}$$

A number of closures to solve this equation have been developed over the years to approximate the exact closure relation

$$h(\mathbf{r})+1=\exp[h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})+B(\mathbf{r})] \tag{32}$$

with $B(\mathbf{r})$ as the so-called bridge function. These approximations are necessary because the bridge function is not known analytically. Just like for the Ornstein-Zernike and the 1D RISM equations there is a similar expression for the hypernetted chain closure where the bridge function is zero at all distances, i.e.

$$h(\mathbf{r})+1=\exp[h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})] \tag{33}$$

This closure performs well given its rather crude approximation of ignoring the bridge function altogether, which is only a good approximation for large distances r. It does, however, not always converge easily, especially for more complex molecular systems and those with strong intermolecular interactions [62]. To avoid these problems, it is also possible to use the partial series expansion closure (PSE-n) that was developed in this group as an extension of the so-called Kovalenko-Hirata (KH) closure, where the closure is approximated with a polynomial equation for values where the exponential might become unstable [72,73,74]. This closure can be written as

$$h(\mathbf{r})+1=\begin{cases}\exp[h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})] & h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})\leq 0\\ \sum_{n}[h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})]^{n}/n! & h(\mathbf{r})-c(\mathbf{r})-\beta u(\mathbf{r})>0\end{cases} \tag{34}$$

where the sum is constructed in such a way that for $n\rightarrow\infty$ the HNC closure is formally recovered, while using small values of n makes it possible to avoid most of the convergence problems associated with the HNC closure. Specifically, the PSE-1 closure is equivalent to the KH-closure.

The primary results of any 3D RISM calculation are thus the total and the direct correlation functions of each interaction site. The former has a very intuitive physical interpretation, it is the solvent site density at each point in space around the solute. Together with the direct correlation function it is also possible to calculate thermodynamic data of the system. For the HNC closure the functional of the excess chemical potential is

$$\mu_{\mathrm{HNC}}^{\mathrm{ex}}=\beta^{-1}\sum_{\gamma}\rho_{\gamma}\int\left(\frac{1}{2}h_{\gamma}{}^{2}(\mathbf{r})-c_{\gamma}(\mathbf{r})-\frac{1}{2}h_{\gamma}(\mathbf{r})c_{\gamma}(\mathbf{r})\right)\mathrm{d}\mathbf{r} \tag{35}$$

while the equivalent expression for the PSE-n closure is

$$\mu_{\mathrm{PSE}\text{-}n}^{\mathrm{ex}}=\mu_{\mathrm{HNC}}^{\mathrm{ex}}-\beta^{-1}\sum_{\gamma}\rho_{\gamma}\int\frac{\Theta(h_{\gamma}(\mathbf{r}))(h_{\gamma}(\mathbf{r})-\beta u_{\gamma}(\mathbf{r})-c_{\gamma}(\mathbf{r}))^{n+1}}{(n+1)!}\mathrm{d}\mathbf{r} \tag{36}$$

where Θ is the Heaviside function [73,75]. The approximations made for these closures lead to a systematic error in the calculated excess chemical potential that will be discussed in more detail in chapter 2.2.

Other thermodynamic data that can be derived from the converged correlation functions is the partial molar volume of the solute in solution. This quantity is the effective change of the volume upon insertion of the particle and thus a measure for the effective size of the solute, including solvent reordering and other effects due to the solute-solvent interactions. The partial molar volume can be calculated for a polyatomic liquid using 3D RISM via either the total or the direct correlation function [76,77]. Derived in analogy to the original Kirkwood-Buff theory the expression for the partial molar volume calculated by using the total correlation function is

$$V_{m,h} = \beta^{-1}\kappa - \int h_\gamma(\mathbf{r})\mathrm{d}\mathbf{r} \tag{37}$$

with κ as the isothermal compressibility [78], whereas for the direct correlation function it can be calculated via

$$V_{m,c} = \beta^{-1}\kappa\left(1 - \rho\sum_\gamma \int c_\gamma(\mathbf{r})\mathrm{d}\mathbf{r}\right) \tag{38}$$

While the experimental isothermal compressibility is used for large parts of this work, both routes only converge to the same result for neutral species if the RISM consistent compressibility is used. This compressibility can be calculated from the reciprocal space 0-element of the direct correlation function using [76]

$$\beta^{-1}\kappa_{\mathrm{RISM}} = \left(\rho\left(1 - \rho\sum_\gamma c_\gamma(k=0)\right)\right)^{-1}. \tag{39}$$

For charged species there is an additional difference between the results calculated from either route, because unlike the partial molar volume of a salt, the partial molar volume of an individual ion at infinite dilution is not thermodynamically well-defined [79,80]. This gives rise to an additional contribution to the partial molar volume so that in practice for ions the partial molar volumes calculated by the different routes differ by

$$V_{m,h} = V_{m,c} - \beta^{-1}\kappa\rho D \tag{40}$$

where D is a solvent-specific constant that has opposite sign for cations and anions [77,79,81]. Experimental techniques that make it possible to actually resolve the individual contributions to the partial molar volume of a salt show that $V_{m,c}$ is the PMV that corresponds to the experimental values [79,82].

2.1.4 EC-RISM

The embedded-cluster reference interaction site model is an application of the 3D RISM solvation model to quantum chemical systems [83]. After an initial calculation of the molecule's wavefunction in a vacuum EC-RISM is an iterative scheme consisting of two operations: First, the electrostatic potential of the solvent molecule resulting from the QC calculation is used as an input for a 3D RISM calculation. The electrostatic potential can be used either in an approximate manner by calculating the atom-wise CHelpG charges, or by using the exact electrostatic potential resulting from the QC calculation [84,85,86].

For polar solvents the solvent distribution resulting from this calculation will lead to an increased polarization of the solute. This effect can be accounted for by generating the charge density around the solute

$$\rho_\gamma^q(\mathbf{r}) = \sum_\gamma q_\gamma \rho_{\gamma,\infty} g_\gamma(\mathbf{r}) \tag{41}$$

using the 3D RISM pair distribution function $g_\gamma(\mathbf{r})$ and the solvent atom's partial charges q_γ with ρ_γ as the bulk density of site γ. This charge density is then discretized onto a grid around the solute which yields point charges

$$q_\gamma(\mathbf{r})\Delta V \tag{42}$$

where ΔV is the grid cell volume, and these can be used as input for a new QC calculation. As mentioned above, an alternative to this approach is to use the exact electrostatic potential derived from the wave function. To achieve a smooth reciprocal space representation of the electrostatic potential the electrostatic interactions are calculated using an Ewald sum approach, where the Coulomb interaction energy calculated from the ESP derived charges

$$u^q(\mathbf{r}_\alpha \mathbf{r}) = \sum_\alpha \sum_\gamma \frac{q_\alpha q_\gamma}{4\pi\varepsilon_0 \mathrm{r}(\mathbf{r}_\alpha \mathbf{r})} \tag{43}$$

with ε_0 as the vacuum permittivity constant is used to represent the long-range interactions. This means that only the difference between the exact and the point-charge based potential needs to be transformed into reciprocal space, and this difference is short-ranged compared to the electrostatic potential itself. Furthermore, these point-charge based interactions do not meaningfully contribute to the evaluation of the interaction energy if the grid is large enough to make the difference between the electrostatic potential energy and the point-charge based electrostatic energy vanish at the edges of the grid [86,87]. However, while working on the SAMPL6 challenge it was noticed that for strongly polar or charged species this renormalization approach can cause convergence problems and inaccurate Gibbs energies, which can be alleviated by using a potential switching approach developed by P. Kibies [87], that effectively scales the electrostatic potential near the edges of the box so that it becomes equal to the potential derived from the point charges.

Iterative application of the EC-RISM cycle of calculating the electronic and the solvent structure until convergence as measured by the change of the EC-RISM Gibbs energy of a molecule described by its atom coordinates $\{\mathbf{r}\}$ in solution

$$G(\{\mathbf{r}\}) \equiv E^{\mathrm{sol}}(\{\mathbf{r}\}) + \mu^{\mathrm{ex}}(\{\mathbf{r}\}) \tag{44}$$

where E^{sol} is the electronic energy resulting from the QC calculation and μ^{ex} is the 3D RISM excess chemical potential yields the solvent-polarized wave function and the solvent distribution functions around the solute. Because the polarized wave function can be further analyzed it is also possible to calculate spectroscopic data of a molecule in solution. The solvent structure around a solute can in principle be used to e.g. place explicit water molecules at the density maxima for additional highly accurate quantum calculations [88,89,90,91]. The reference state in solution for all EC-RISM calculations is assumed to have a pressure of 1 bar, a formal concentration of 1 M at a temperature of 298.15 K and infinite dilution conditions.

2.2 Calculation of physicochemical properties

2.2.1 Empirical correction of the excess chemical potential

One well-known deficiency of 3D RISM theory is the inability to give accurate results when calculating solvation free energies due to an overestimation of the pressure leading to an overestimation of the energy required to form a solute-sized cavity in the solvent [14,15,16]. Since this error should be highly correlated to the size of the molecule a correction using the partial molar volume of the molecule is commonly applied to correct for this deficiency in the calculated excess chemical potential μ^{ex} [92,93]. For example, the universal correction (UC) developed by Palmer *et. al* corrects the RISM excess chemical potential using

$$\mu^{\mathrm{UC}} = \mu^{\mathrm{ex}} + aV^{m} + b \tag{45}$$

where V^m is the partial molar volume and a and b are parametrized by minimizing the difference between μ^{UC} and experimental Gibbs energies of hydration. Similar approaches were developed by Truchon *et al*. where they scale the direct correlation functions to to correct for the bridge function inside the solute volume [94,95], or the PC/PC+ pressure correction developed by Sergiievskyi *et al.* [96]. Such models however had not been used in combination with the quantum-chemical embedding approach of EC-RISM until D. Tomazic developed and tested a large number of corrections [44]. In their most general form, these corrections can be written as

$$\mu^{\mathrm{ex}}_{\mathrm{corr}} = c_{\mu}\mu^{\mathrm{ex}} + c_{v}V^{m} + c_{q}q + c_{o} \tag{46}$$

where q is the solute charge, the parameter c_v is the basis of the partial molar volume correction while the additional parameter c_q is necessary to correct for between the physical (experimental) process of ion solvation and the unphysical process modeled by 3D RISM [1,16]. In the former, the solute crosses an explicit vacuum-water interface during the solvation process, while in RISM theory the solvent is for all intents and purposes “infinite”, i.e. it has no surface [97,98]. The surface polarization then gives rise to an additional term in the solvation free energy that has to be explicitly accounted for when predicting Gibbs energies of solvation or other properties [99,100,101]. As shown in eq. (40) there is an additional additive contribution to the PMV resulting from a nonzero net charge that might in principle correct for this discrepancy, but it is too small to yield accurate Gibbs energies without the additional parameter.

The two free parameters c_μ and c_o differ from the other two parameters in that they have no sound theoretical basis which means that they should ideally be 1 and 0, respectively. In some cases, it can still be beneficial to use these additional parameters to correct for approximations in e.g. the solvent description or the force field used for the intermolecular interactions, but care must be taken that no overfitting occurs. It should also be noted that there is no universal correction applicable to every EC-RISM calculation, but instead the best combination of parameters differs slightly between different QC levels of theory, basis sets and closures, and there can be significant deviations between the parameters for different solvents.

Regardless of which set of parameters is used, the correction must be trained with experimental data. In this and in previous works the Gibbs energy of solvation is used as reference data as it can be easily calculated using EC-RISM (or 3D RISM by ignoring the electronic contributions to the total energy) via

$$\Delta_{\mathrm{solv}} G = E^{\mathrm{sol}} + \mu^{\mathrm{ex}} - E^{\mathrm{vac}}, \tag{47}$$

where E^{sol} is the electronic energy of the solute in solution while E^{vac} is the electronic energy of the solute in the gas phase. This approximation only accounts for vibrational and rotational contributions to the Gibbs energy by effectively parametrizing them into the PMV correction by fitting to experimental data, but these can in principle be explicitly taken into account using e.g. the rigid rotor, harmonic oscillator (RHHO) model [3]. Further improvements to the accuracy of the calculated Gibbs energy of solvation can be made using high-level coupled cluster energies for the gas phase [102]. This quantity is suitable to use for the parametrization because unlike the excess chemical potential it is experimentally accessible and the error in the electronic energies derived from first principles should be insignificant compared to the error in the excess chemical potential.

2.2.2 Calculation of acidity constants

In the most general form, the acid dissociation constant K_{a} of a molecule is the equilibrium constant of the formal dissociation reaction

$$\mathrm{HA}_{\mathrm{aq}} \rightarrow \mathrm{H}^{+}_{\mathrm{aq}} + \mathrm{A}^{-}_{\mathrm{aq}} \tag{48}$$

where the charge of the species $\mathrm{HA}_{\mathrm{aq}}$ is not restricted to be 0, but for a charged $\mathrm{HA}_{\mathrm{aq}}$ the charge of the $\mathrm{A}^{-}_{\mathrm{aq}}$ changes accordingly. The acidity constant can then be defined as

$$-\beta \ln K_{\mathrm{a}} = \Delta_r G^0 = -\beta \ln \frac{a\left(\mathrm{H}^{+}_{\mathrm{aq}}\right) a\left(\mathrm{A}^{-}_{\mathrm{aq}}\right)}{a\left(\mathrm{HA}_{\mathrm{aq}}\right)} = \mu^0\left(\mathrm{H}^{+}_{\mathrm{aq}}\right) + \mu^0\left(\mathrm{A}^{-}_{\mathrm{aq}}\right) - \mu^0\left(\mathrm{HA}_{\mathrm{aq}}\right) \tag{49}$$

where a denotes the activities of the respective species, $\Delta_r G^0$ the standard Gibbs energy of reaction, and μ^0 the standard chemical potentials [3]. To calculate acidity constants resulting from multiple distinct states such as conformational states or tautomers of the same molecule, two main approaches can be evaluated. Under the assumption that the standard chemical potentials can be approximated as the sum of an ideal contribution and the EC-RISM Gibbs energy defined above, the reaction free energy can be calculated using a canonical partition function of the form

$$\begin{aligned}\Delta_r G^0 &= \mu^{0,\mathrm{id}}\left(\mathrm{H}^{+}_{\mathrm{aq}}\right) + \mu^{0,\mathrm{id}}\left(\mathrm{A}^{-}_{\mathrm{aq}}\right) - \mu^{0,\mathrm{id}}\left(\mathrm{HA}_{\mathrm{aq}}\right) \\ &+ \mu^{\mathrm{ex}}\left(\mathrm{H}^{+}_{\mathrm{aq}}\right) - \beta^{-1} \ln \frac{\sum_{j=1}^{M} \exp\left[-\beta G_j\left(\mathrm{A}^{-}_{\mathrm{aq}}\right)\right]}{\sum_{k=1}^{N} \exp\left[-\beta G_k\left(\mathrm{HA}_{\mathrm{aq}}\right)\right]}\end{aligned} \tag{50}$$

where “id” and “ex” denote the ideal and the excess parts of the chemical potential, respectively, M the number of base states, and N the number of acid states. The Gibbs energy of the proton cannot be calculated using standard EC-RISM approaches but it is possible to sum it with the ideal contributions to an additive constant b. An in-depth discussion of the dependence of this “constant” on molecular parameters can be found in ref. 3, but this variation is generally negligible. The expression then becomes

$$\mathrm{p}K_{\mathrm{a}}^{\mathrm{PF}}=\frac{\beta\Delta_{\mathrm{r}}G^{0}}{\ln 10}=b-\frac{m}{\ln 10}\ln\frac{\sum_{j=1}^{M}\exp\left[-\beta G_{j}\left(\mathrm{A}_{\mathrm{aq}}^{-}\right)\right]}{\sum_{k=1}^{N}\exp\left[-\beta G_{k}\left(\mathrm{HA}_{\mathrm{aq}}\right)\right]} \tag{51}$$

where b contains the Gibbs energy of hydration of the proton and the ideal terms of the excess chemical potentials, and m is an additional parameter that is ideally 1 [103,104,105]. These parameters can be fitted with experimental data to achieve accurate predictions of acidity constants [106,107].

An alternative approach is the state transition approach, hereafter denoted as “ST”, where all states where the deprotonated state is the result of a single deprotonation step with no further tautomerization are connected as individual reactions

$$\mathrm{HA}_{\mathrm{aq},k}\rightarrow\mathrm{H}_{\mathrm{aq}}^{+}+\mathrm{A}_{\mathrm{aq},j}^{-} \tag{52}$$

The equivalent expression to eq. (51) is

$$\mathrm{p}K_{\mathrm{a},jk}=b+\frac{m\beta}{\ln 10}\left[G_{j}\left(\mathrm{A}_{\mathrm{aq}}^{-}\right)-G_{k}\left(\mathrm{HA}_{\mathrm{aq}}\right)\right] \tag{53}$$

and as shown by Bochevarov et al. there is a closed form expression for the macrostate acidity constant resulting from the individual microstate acidity constants [108], namely

$$K_{\mathrm{a}}^{\mathrm{ST}}=\sum_{j=1}^{M}\frac{1}{\sum_{k=1}^{N}\frac{1}{K_{a,jk}}} \tag{54}$$

These approaches yield the same results only if the slope parameter $m=1$, because when applying the corrections, the acidity constant for the two approaches is calculated as

$$\begin{aligned}K_{\mathrm{a}}^{\mathrm{PF}}&=10^{-b}\left(\frac{\sum_{j=1}^{M}\exp\left[-\beta G_{j}\left(\mathrm{A}_{\mathrm{aq}}^{-}\right)\right]}{\sum_{k=1}^{N}\exp\left[-\beta G_{k}\left(\mathrm{HA}_{\mathrm{aq}}\right)\right]}\right)^{m}\\&\neq 10^{-b}\frac{\sum_{j=1}^{M}\exp\left[-\beta G_{j}\left(\mathrm{A}_{\mathrm{aq}}^{-}\right)\right]^{m}}{\sum_{k=1}^{N}\exp\left[-\beta G_{k}\left(\mathrm{HA}_{\mathrm{aq}}\right)\right]^{m}}=K_{\mathrm{a}}^{\mathrm{ST}}.\end{aligned} \tag{55}$$

Only if there is only one acid and one base microstate equality is conserved even for non-zero slope parameters. While the deviations are minor for usual sizes of *m* found for EC-RISM this cannot be guaranteed for all methods for predicting acidity constants from first principles and the choice of using the partition function or the state transition approach may have statistically significant effects on the predictions [3].

Given the pK_a as defined above L. Eberlein was able to show that it is possible to calculate the ionization state fractions for any number of titratable sites *n* of which *i* are deprotonated [2,3,109]. This corresponds to the deprotonation reaction $\mathrm{AH}_{n-i} \rightarrow \mathrm{AH}_{n-i-1} + \mathrm{H}^{+}$ with the equilibrium constant

$$K_{a,i+1} = \frac{a_{+}\, a_{i-1}}{a_i} = \frac{a_{+}\, x_{i-1}}{x_i} \tag{56}$$

under the assumption that the activities *a* can be replaced by the molar fractions *x*. By calculating the individual fractions using Henderson-Hasselbalch type equations and extrapolating the results to *n* titratable sites the fraction of an ionization state is

$$x_i = (10^{-\mathrm{pH}})^{-i} \prod_{j=0}^{i} 10^{-\mathrm{p}K\mathrm{a},j} \left(1 + \sum_{k=1}^{n} (10^{-\mathrm{pH}})^{-k} \prod_{l=1}^{k} 10^{-\mathrm{p}K\mathrm{a},l} \right)^{-1} \tag{57}$$

where the activity of the proton was replaced by $10^{-\mathrm{pH}}$ and an empty product is assumed to be unity [105]. The fractional, pH-dependent tautomer populations can then be calculated as

$$x_{it} = x_{it|i} x_i \tag{58}$$

with the conditional tautomer population

$$x_{it|i} = \exp(-G_{it} / RT) / \sum_{t} \exp(-G_{it} / RT)\,. \tag{59}$$

2.2.3 Partition and distribution coefficients

The calculation of neutral state partition coefficients between two immiscible phases is more straightforward. Given an aqueous phase “w” and an organic phase “o” the partition coefficient of a molecule is given by

$$\log P = -\frac{\Delta_{\mathrm{trans}} G^{0}}{RT \ln 10} = \frac{G^{0}_{\mathrm{wat}} - G^{0}_{\mathrm{oct}}}{RT \ln 10}\,. \tag{60}$$

This equation shows that the Gibbs energies in the respective solvents must be very accurate, as an error of 1 kcal/mol in the transfer free energy between the solvents will lead to an error of approximately 0.73 in the log *P* and the dynamic range of the partition coefficient is significantly smaller than that of e.g. solvation free energies.

For the distribution coefficient log *D*, a more complicated picture arises [110,111]. Under the assumption that only the neutral species enters the organic phase, the log *D* is calculated as

$$\log D = \log P - \log\left(1 + 10^{\mathrm{p}K_\mathrm{a} - 7.4}\right) \tag{61}$$

for bases and

$$\log D = \log P - \log\left(1 + 10^{\mathrm{p}K_\mathrm{a} + 7.4}\right) \tag{62}$$

for acids. Depending on the organic solvent under investigation this assumption does not always hold, and more complicated methods must be used. In that case the distribution coefficient of a base is

$$\log D = \log\left(P + P_i \cdot 10^{(\mathrm{p}K_a - \mathrm{pH})}\right) - \log\left(1 + 10^{(\mathrm{p}K_a - \mathrm{pH})}\right) \tag{63}$$

for bases and

$$\log D = \log\left(P + P_i \cdot 10^{(\mathrm{pH} - \mathrm{p}K_a)}\right) - \log\left(1 + 10^{(\mathrm{pH} - \mathrm{p}K_a)}\right) \tag{64}$$

for acids where P_i is the partition coefficient of the ionic species. In either case, no unified formula for acids and bases exists because the definition of the acid dissociation constant determines the direction of the protonation reaction.

In this work acids and bases are defined not by their behavior in aqueous solution, where compounds that are proton donors at a pH of 7 are acids and proton acceptors are bases, but more generally. Any neutral or negatively charged compound and its deprotonated form are defined as an acid and its conjugate base, regardless of the pH that is necessary to deprotonate it. Conversely, any neutral or positively charged compound and its protonated form are considered to be a base and its conjugate acid. This definition is in line with the Brønsted-Lowry acid-base theory and is more robust for modeling purposes, even if it is not necessarily in line with the common understanding of acids and bases

2.3 Molecular dynamics simulations

2.3.1 Introduction

Molecular dynamics (MD) simulations have been used for decades to investigate molecular systems at an atomic level. The major advantage of MD simulations when compared to quantum mechanical (QM) methods is that much larger systems consisting of up to millions of atoms can be investigated [112]. MD simulations work by propagating the atomic movements of the system through time. Given a good parametrization and long enough sampling this should in theory yield the equilibrium distribution of states that the real system occupies [113].

The interatomic potentials used in MD simulations are twofold. Firstly, the long-range potentials capture the effect of London dispersion, Pauli repulsion, and Coulomb interactions through the charge q and usually through the Lennard-Jones parameters ε and σ. For atoms that are part of the same molecule there are additional terms to model bonds and angles, usually with simple quadratic potential functions, and dihedrals, for which trigonometric functions are used to allow for a free rotation of the bonds. While atoms that are more than three bonds apart are generally treated as though they were in different molecules, the van der Waals and electrostatic interactions between atoms separated by only one or two bonds are usually switched off to avoid numerical issues. For atoms separated by exactly three bonds the so-called 1-4 interactions may be scaled down by a constant factor to account for the partial inclusion of the short-ranged interactions in the torsional potentials [114]. Especially for systems that are aromatic or have a similarly inhibited free dihedral rotation so-called improper torsions are used to prevent this. The sum of all these contributions

$$U_{\text{tot}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}} + U_{\text{coulomb}} + U_{\text{LJ}} \tag{65}$$

the total potential U_{tot}, consists of the bonded terms U_{bond}, U_{angle}, and U_{torsion} that represent the molecular bonds and the nonbonded terms U_{coulomb} and U_{LJ} that represent the electrostatic and the repulsive-dispersive interaction, respectively. From this the acceleration of a particle can be calculated by taking the gradient and dividing by its mass m

$$\mathbf{a}_i = \mathbf{F}_i / m = -\nabla U(\mathbf{r}_i) / m_i. \tag{66}$$

Here $\mathbf{a}_i$ is the vector of acceleration, $\mathbf{F}_i$ the vector of the force, and $\mathbf{r}_i$ the position of the particle denoted as i, respectively. Propagation through time can be achieved by using e.g. the velocity Verlet algorithm [115], where the positions and velocities of the atoms for the next timestep are calculated as

$$\mathbf{r}_i(t+\delta t)=\mathbf{r}_i(t)+\delta t\cdot\mathbf{v}_i(t)+\frac{1}{2}\delta t^2\cdot\mathbf{a}_i(t) \tag{67}$$

and

$$\mathbf{v}_i(t+\delta t)=\mathbf{v}_i(t)+\frac{1}{2}\delta t^2\cdot(\mathbf{a}_i(t)+\mathbf{a}_i(t+\delta t)) \tag{68}$$

where $\mathbf{r}_i$ and $\mathbf{v}_i$ are the positions and velocities of the atoms, respectively. t is the current time and δt is the discrete timestep with which the system is propagated. The timestep has to be chosen large enough that sufficient sampling can be achieved in a given simulation, but if it is chosen too large errors will accumulate, especially if the timestep is larger than the fastest motion in the system. These are usually the vibrations of hydrogen atoms and by holding them fixed a timestep of 2 fs is possible [116].

2.3.2 Liquid structure from MD simulations

Among the information that can be extracted from MD simulations, structural information is the primary one. A trajectory generated by an MD simulation consists of the atom positions and velocities for each time step that is saved. The systems investigated by MD simulation are generally assumed to be ergodic and so the equilibrium distribution functions of the system arise naturally given a long enough simulation time. For RISM, as shown in greater detail in Chapter 2.1.2, the radial distribution function $g(r)$ is of great interest, because it can be used to generate solvent models. The radial distribution function describes the number of particles that can be found in a distance r, relative to the bulk density. Thus, it can be used as a graphical representation of the depletion and enrichment of solvent atoms around a particle due to the molecular interactions as shown in Fig. 1.

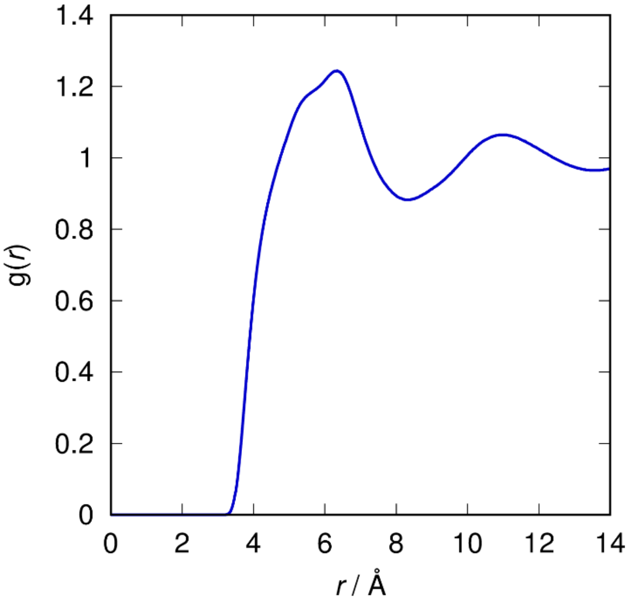


Fig. 1: Radial distribution function between united atom cyclohexane CH_2 groups extracted from the molecular dynamics simulation described in chapter 3.3.1.

To understand how to generate inter- and intramolecular distribution functions from MD simulations to use for RISM it is convenient to consider the model system of pure water. The hydrogen atoms in water are positioned symmetrically around the central oxygen atom which makes water part of the C_{2v} point group. Due to this symmetry the two hydrogen atoms are equivalent, so for the purpose of generating distribution functions only one type of hydrogen atom exists. This model system yields three different distribution functions: g_{OO}, g_{HH}, and g_{OH}, the latter of which is identical to a hypothetical g_{HO} function. The intermolecular distribution functions can be discretized from the simulation data via

$$g(r) = \frac{N(r \pm \Delta r / 2)}{V(r \pm \Delta r / 2)\rho} \tag{69}$$

where $N(r \pm \Delta r / 2)$ is the number of atoms of a certain type found in the interval $r \pm \Delta r / 2$ and $V(r \pm \Delta r / 2)$ the volume of the same interval. The g_{OO} function would be generated by applying eq. (69) from each oxygen atom and counting all other oxygen atoms, the g_{HH} function by applying it from each hydrogen atom and counting all hydrogen atoms that are not part of the same molecule and so on. To generate the intramolecular distribution functions the same scheme is applied but, in this case, counting only those atoms that are part of the same molecule. Since the hydrogen atoms of water are usually kept frozen to enable larger step sizes this is only necessary for larger molecules with more conformational degrees of freedom.

2.3.3 Thermodynamic integration

MD simulations are not only used to gain structural information of a molecular system but also thermodynamic properties. One method to calculate Gibbs energies of solvation that are of great interest in this work is thermodynamic integration (TI). This method makes it possible to calculate the difference in the Helmholtz energy between two pre-defined states by connecting them with an integration variable λ that scales the interactions of the alchemically active region with the rest of the system. The partition function and the Helmholtz energy *A* that can be calculated from it are then functions of the scaling parameter λ

$$A(\lambda) = -\beta^{-1} \ln Z(\lambda) \tag{70}$$

Here *Z* is the classical canonical ensemble partition function and $\beta = 1/(k_B T)$. Differentiating with respect to the integration variable gives [117]

$$\frac{\partial A(\lambda)}{\partial \lambda} = \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{71}$$

where *H* is the system Hamiltonian and the angled brackets imply a canonical average that is computed from the MD simulation [118]. From this the difference in the Helmholtz energy between the states defined by $\lambda = 0$ and $\lambda = 1$ can be calculated by integrating

$$\Delta A = A(1) - A(0) = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{72}$$

This integration cannot be done analytically, but instead a number of λ-states are simulated and numerical integration yields the difference in the Helmholtz energy. Depending on the system setup and the choice of the alchemically active region it is possible to calculate properties such as (relative) binding free energies and solvation free energies. Some care must be taken, because in the limit of $\lambda = 0$ the potential energy becomes infinite for the standard potential energy functions used in MD simulations. Usually, soft-core potentials are used to avoid this issue [119,120].

3 Generation of solvent susceptibilities

3.1 Introduction

3.1.1 Cyclohexane

The SAMPL5 challenge on predicting cyclohexane-water distribution coefficients necessitated the development of a cyclohexane solvent susceptibility for use in 3D RISM. The cyclohexane solvent model developed in this work is based on the united atom GROMOS96 45A3 parameter set for aliphatic hydrocarbons, because this reduces the number of interaction sites from 18 to 6 and unlike in the case of water the hydrogen atoms are not involved in directional polar interactions [124]. Each interaction site is thus a sphere representing an entire CH_2 group instead of individual atoms.

Due to certain issues encountered while generating the solvent susceptibilities using 1D RISM that will be expanded upon below it was necessary to use MD simulations to generate them. The solvent susceptibilities were generated by measuring the likelihood of finding a solvent atom of either the same or a different solvent molecule in a certain distance through histogram analysis. This had the additional benefit of making it possible to fully resolve the intramolecular distribution function that is commonly represented through the Dirac delta function, implying a single, fully rigid conformation.

3.1.2 Octanol and water-octanol mixtures

A model for neat octanol had already been developed and applied in EC-RISM calculations by the author [121]. In this work a new model for water-saturated octanol was developed to account for the experimental conditions usually found in experiments for partition or distribution coefficient determinations. When using a shake-flask protocol the two phases are in direct contact and thus a certain amount of octanol is dissolved in the water phase and a significantly larger amount, up to 48.91 g·kg^{-1} of water is dissolved in the octanol phase [122]. In theory it would be necessary to generate an octanol-saturated water model as well, but the amount of octanol dissolved in water is significantly lower at only 0.3 g·L^{-1}. Furthermore, while the water content of the octanol phase should increase the polar character of this phase considerably the small amount of octanol in the water phase is not expected to have such a significant effect on the solvent properties.

3.2 Computational details

3.2.1 Cyclohexane MD simulation

To set up the simulation packmol 1.1.2.023 was used to place 10,000 cyclohexane molecules in a cubic box with an edge length of 122 Å [123]. The molecules were parametrized using the united atom GROMOS96 45A3 parameter set for aliphatic hydrocarbons [124], leading to a total system size of 60,000 atoms. For the simulation Berendsen temperature and pressure controls were used to keep the system at 298.15 K and maintain a pressure of 1 bar. The time constants for the temperature and pressure were 0.1 ps and 2.5 ps, respectively, and the experimental compressibility of 1.12·10^{-4} bar^{-1} was used [125]. As in the original work on the GROMOS96 45A3 force field all bonds were kept constrained using the implementation of the SHAKE algorithm implemented in Gromacs [126,127]. For the Lennard-Jones interactions a cutoff of 1.4 Å was used, while the Coulomb cutoff could be ignored since every atom of the system is electrostatically neutral.

Using Gromacs 4.6.3 the system was first minimized until the greatest force evaluated converged to below 750 kJ mol^{-1} nm^{-1} and then simulated for 17 ns using a time step of 2 fs. Every 5 ps, i.e. 2500 time steps, a snapshot of the system was taken, and the final 15 ns of this continuous trajectory were used to extract the intra- and intermolecular distribution functions.

Due to the extended functionality of the more modern version of the tool, the intramolecular distribution functions were extracted using the “gmx_d distance” tool of Gromacs 5.0 with a bin width of 0.002 Å. This was achieved by creating custom index files containing only the pairs of atoms that are part of the same molecule. For each intramolecular distribution function (e.g. the one between C1 and C2, C1 and C3, etc.) a new index file had to be created, leading to a total number of 15 index files.

For the intermolecular distribution functions the version of the “g_rdf” tool found in Gromacs 4.6.3 was sufficient and used accordingly. By excluding atoms within the same molecule only the intermolecular distribution functions were extracted from the trajectory with this tool.

3.2.2 Solvent susceptibilities and 1D RISM calculations

The solvent susceptibilities for neat octanol had already been generated in a previous work with the same settings used for the octanol-water mixtures described below in Table 2, but using a number density of $3.82054 \cdot 10^{-3}$ Å^{-3} and the experimental dielectric permittivity of 9.86284 [121,128]. The molecular structure and solvent parameters for both cyclohexane and octanol are also provided as OR_01. The structure for cyclohexane represents the average distances between two sites during simulation, while for octanol it is the basis for the generation of rigid intramolecular distribution functions as described in equation (28). To generate solvent susceptibilities for octanol-water mixtures, SPC/E water sites were added to the 1D RISM calculations, using number densities of $1.37473 \cdot 10^{-3}$ Å^{-3} for the water sites and $3.64253 \cdot 10^{-3}$ Å^{-3} for the octanol sites. Furthermore, compared to neat octanol with a dielectric permittivity of 9.86294, for the octanol-water mixtures a lower permittivity of 9.1 was used. These values are slightly inaccurate because they were estimated from the molar mass-scaled saturation fraction of wet *n*-octanol of 0.274 and its mass density of 0.82883 [129]. Future works should use the accurate number densities of $1.3598 \cdot 10^{-3}$ Å^{-3} for the water sites and $3.65787 \cdot 10^{-3}$ Å^{-3} for the octanol sites, as calculated from the experimental densities and a saturation water mole fraction of 0.2705 [122].

Since no experimental dielectric permittivity for water-saturated octanol at a temperature of 298.15 K was available the permittivity was estimated by extrapolating the available data at 293.15 K and 303.15 K to the saturation mole fraction and adding the average offset from neat octanol at those temperatures to the experimental value at 298.15 K. The resulting permittivity for *n*-octanol that should be used in future works is 8.41 [130].

All 1D RISM calculations for generating the octanol solvent susceptibilities were conducted using the same logarithmic grid as for the cyclohexane model. The RISM integral equations were solved using a Mathematica implementation developed by Professor Stefan M. Kast until the maximum norm of the direct correlation functions was smaller than 10^{-6}.

3.3 Results and discussion

3.3.1 Cyclohexane

Initially, it was attempted to use 1D RISM to generate solvent susceptibilities for use with EC-RISM, but the appearance of singularities in the generated susceptibilities made this approach impossible. The relevant parameters used for the generation of the solvent susceptibility are summarized in Table 1, σ and ε are the Lennard-Jones parameters of the cyclohexane CH_2 groups, ρ_{cyc} the experimental density and ε_r the dielectric constant.

Table 1: Solvent parameters for the generated susceptibility for cyclohexane [124,128].

	Cyclohexane
$\sigma(CH_2)$ /Å	3.95474
$\varepsilon(CH_2)$ /10^{-21} J	0.795858
ρ_{cyc} /10^{-3} Å^{-3}	5.488
ε_r	2.01647

For this reason, the choice was made to extract the intra- and intermolecular distribution functions from MD simulations. This is not as straightforward as it might seem, because usually the simulated total correlation functions are switched to the analytic HNC total correlation functions [67,69], because the minimum image convention restricts the maximum length of the former, and a prohibitively large simulation box would be needed to be able to apply the usual logarithmic grid ranging from 0.0059 to 164.02 Å [131,132].

Here, the intramolecular functions were approximated using a single Gaussian function of the form

$$\omega(r) = a \cdot \exp\left(-\frac{(r-b)^2}{c^2} \right) \tag{73}$$

with variable height a, position b and width c to give an analytical representation in reciprocal space. The Gaussian fit parameters are gathered in OR_01. The intermolecular distribution functions were smoothed before calculation of the solvent susceptibility using cubic splines, where the smoothing factor S was calculated from the number of molecules n_{mol} and the number of snapshots n_t of the system [69,133] via the equation

$$S = \frac{n_{mol} \cdot n_t}{2}. \tag{74}$$

Beyond a distance of 61.19 Å the intermolecular distribution function was set to one, interpolated on a logarithmically spaced grid of 512 points ranging from 0.0059 to 164.02 Å and

and after calculating the solvent susceptibilities using eq. (27) they were converted to reciprocal space using a numerical Fourier transformation.

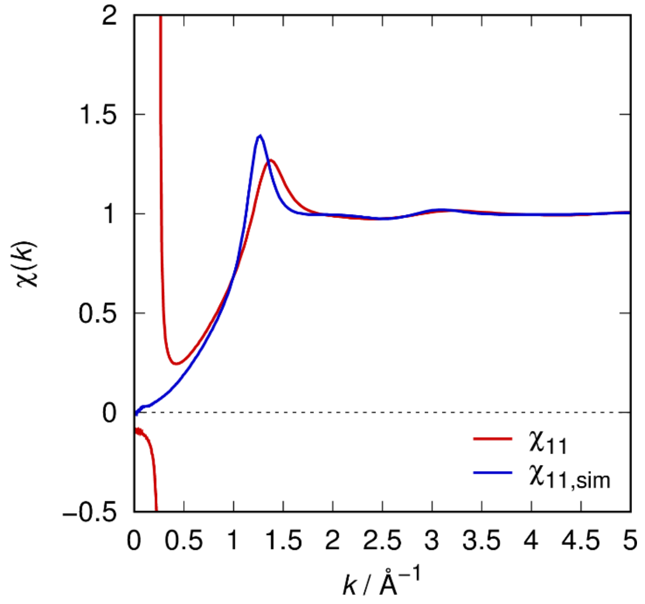


Fig. 2: Comparison of solvent susceptibilities for cyclohexane generated with 1D RISM (red) and from direct inversion of simulated correlation functions (blue).

As can be seen in Fig. 2, the discontinuity in the solvent susceptibility vanishes when using the distribution functions generated from the MD simulation while the overall shape and the behavior towards the limits stays similar to that found in the discontinuous solvent susceptibility function derived from 1D RISM. This approach made it possible to use cyclohexane as a solvent in 3D RISM calculations for the SAMPL5 challenge. It is noteworthy that there is technically no need to use the EC-RISM framework for calculations with this solvent because all methylene groups have a charge of zero and no solute-solvent polarization occurs. Still, it was advantageous to be able to use the same workflow for setting up, running, and analyzing the calculations on the same compounds in different solvents and this was done throughout this work.

3.3.2 Wet *n*-octanol

There were no convergence issues while generating the solvent susceptibilities for wet octanol, but some remarks on the differences between dry and wet octanol, as well as on the differences between the two combinations of density and permittivity are in order. The original number densities were calculated from slightly older experimental values and in a slightly inaccurate manner that ultimately results in the difference of about 1.1% for the water densities and 0.4% for the octanol densities. Additionally, the estimated dielectric constant of water-saturated octanol at 25°C is significantly lower than the one used originally as well. Fortunately, the greatest change in the partition coefficients calculated using the two different solvent susceptibilities is in the order of 0.01 pK units and most are even lower than that. This means that the results generated from the original solvent susceptibilities are close to identical to the ones generated with the new ones. For this reason, while in the future the updated susceptibilities should be used to be as close to the experimental conditions as possible, all data published in this work has been generated with the old susceptibilities. An overview over the different combinations of solvent parameters is given in Table 2. Here σ and ε are the Lennard-Jones parameters of the different chemical groups of the octanol molecule, ρ_{oct} and ρ_{wat} the density used for the octanol and water sites, respectively, and ε_{r} the dielectric constant.

Table 2: Solvent parameters for the generated susceptibilities for octanol and octanol-water mixtures [129, 130,128]. CH_{2o} denotes the CH_2 group next to the alcohol oxygen, specifically.

	Octanol (dry)	Octanol (wet)	Octanol (wet) corrected
$\sigma(CH_3)$ /Å	3.9048	3.9048	3.9048
$\varepsilon(CH_3)$ /10^{-21} J	1.0432	1.0432	1.0432
$q(CH_3)$ /e	0.0000	0.0000	0.0000
$\sigma(CH_2)$ /Å	3.9048	3.9048	3.9048
$\varepsilon(CH_2)$ /10^{-21} J	0.8206	0.8206	0.8206
$q(CH_2)$ /e	0.0000	0.0000	0.0000
$\sigma(CH_{2o})$ /Å	3.9048	3.9048	3.9048
$\varepsilon(CH_{2o})$ /10^{-21} J	0.8206	0.8206	0.8206
$q(CH_{2o})$ /e	0.2650	0.2650	0.2650
$\sigma(O_{\text{oct}})$ /Å	3.0700	3.0700	3.0700
$\varepsilon(O_{\text{oct}})$ /10^{-21} J	1.1823	1.1823	1.1823
$q(O_{\text{oct}})$ /e	-0.7000	-0.7000	-0.7000
$\sigma(H_{\text{oct}})$ /Å	1.0000	1.0000	1.0000
$\varepsilon(H_{\text{oct}})$ /10^{-21} J	0.395589	0.395589	0.395589
$q(H_{\text{oct}})$ /e	0.4350	0.4350	0.4350
$\sigma(O_{\text{wat}})$ /Å	-	3.1660	3.1660
$\varepsilon(O_{\text{wat}})$ /10^{-21} J	-	1.0797	1.0797
$q(O_{\text{wat}})$ /e	-	-0.8476	-0.8476
$\sigma(H_{\text{wat}})$ /Å	-	1.0000	1.0000
$\varepsilon(H_{\text{wat}})$ /10^{-21} J	-	0.3891	0.3891
$q(H_{\text{wat}})$ /e	-	0.4238	0.4238
$\rho_{\text{oct}}/10^{-3}$ Å^{-3}	3.82054	3.64253	3.65787
$\rho_{\text{wat}}/10^{-3}$ Å^{-3}	-	1.37473	1.3598
ε_{r}	9.86294	9.1	8.41

4 SAMPL5: Calculation of cyclohexane–water distribution coefficients log $D_{7.4}$

4.1 Introduction

The SAMPL series of challenges traditionally had the task of predicting hydration free energies of either diverse sets of organic molecules or series of related compounds to test the various solvent models on both overall performance and potential systematic errors. The SAMPL5 challenge changed this pattern for the first time and instead tasked the participants with predicting the distribution coefficients of 53 molecules between cyclohexane and water at a pH of 7.4, which is generally considered to be the physiological pH of the human body. These compounds were further divided into three batches of 13, 20, and 20 molecules respectively, chosen in such a way that the dynamic range of each batch, i.e. the range from the lowest to the highest log $D_{7.4}$ is approximately the same in each batch, and the size of the compounds increases from batch 0 to batch 2. Since the different batches were generated before the experimental results were available the dynamic range was estimated using an empirical log P prediction [51].

As described in chapter 2.2.3 one part of the distribution coefficient of a compound is the neutral state partition coefficient. To calculate this, it was necessary to generate a PMV correction for both water and cyclohexane. Furthermore, the pK_a of the molecule had to be determined to account for the existence of ionic species that cannot enter the organic phase. Due to the time limit imposed by the challenge authors, here, only the distribution coefficients of the 33 molecules in the sets “batch 0” and batch 1” could be submitted in the challenge. To better showcase the development of this work, the data in this chapter is a subset of the results shown in chapter 6.4, where “batch 2” results are included and more advanced solvent and pK_a models are used. The original results of the SAMPL5 challenge are shown here to better demonstrate the advancements that were made alongside and often inspired by problems posed during the SAMPL challenges. While the partition coefficient could be calculated directly from the EC-RISM Gibbs energies of the compound in the respective solvent according to eq. (60), in this case it was calculated from the Gibbs energies of solvation. In theory this should not lead to different results, as the vacuum energies cancel, but here the vacuum conformations for the calculation of the Gibbs energy of solvation in water were reoptimized from the water optimized PCM structures, and vice versa for the Gibbs energy of solvation in cyclohexane. This can give rise to an artificial reorganization term upon calculation of the partition coefficient if the vacuum structures do not converge to the same local minimum. Furthermore, at the time of the challenge deadline there was no model to predict pK_a values yet and the acidity constants were predicted using MoKa and Corina [134,135]. The EC-RISM pK_a model was generated in the post-submission phase and made the prediction of distribution coefficients using only EC-RISM possible.

4.2 Computational details

In analogy to the earlier work by D. Tomazic, this work uses the Minnesota Solvation Database (MNSOL) as reference data to train the partial molar volume correction. This database contains Gibbs energies of solvation for molecules in water and a variety of organic solvents, including *n*-octanol and cyclohexane.

To account for the potentially high conformational flexibility of molecules with a high number of rotatable bonds the same workflow was used for generating the conformations of the molecules in the MNSOL and the SAMPL5 dataset for each solvent: First, for each molecule containing less than 7 rotatable bonds 50 conformations were generated, while for molecules containing more than 7 rotatable bonds 200 conformations were generated using the EmbedMultipleConfs function of RDKit [136,137]. These conformations were then preoptimized using antechamber from the Amber12 software package with an ALPB water model using the dielectric constant corresponding to the respective solvents to account for solvation effects, AM1-BCC charges and GAFF version 1.7 parameters for the non-bonded terms, which are identical to the GAFF version 1.4 parameters and the versions in between

[138,139,140,141]. The resulting structures were clustered based on the following criteria: all conformations with a molecular mechanics (MM) energy at least 20 kcal/mol higher compared to the lowest energy conformation found were discarded. The minimum structure was then assigned as the first cluster and the root mean square distance (RMSD) of the next best structure was determined using the GetBestRMS function of RDKit. If this structure had an RMSD of less than 0.5 Å, the structure was assumed to be properly represented by the existing cluster. If the RMSD was greater, the structure was instead assigned as a new cluster against which all further conformations were compared as well.

All cluster representatives generated that way were then optimized quantum-chemically at the IEFPCM/B3LYP/6-311+G(d,p) level of theory using Gaussian 09 [142] and clustered again using the same workflow as used for the MM structures. Of those cluster representatives, up to five conformations with the lowest PCM energy were used to calculate the solvation free energy using EC-RISM. The vacuum conformations were generated by reoptimizing these structures without a PCM solvent model. No attempt was made to control for shifts in the tautomeric state of the molecules during QM optimization under the assumption that any such shift leads to a lower energy tautomer and only the energetically lowest states contribute to the final Gibbs energy.

To calculate the pK_a values for the Klicić dataset the rotamers and tautomers were optimized at the same level of theory and their energies calculated with EC-RISM [143]. To calculate the pK_a values of the SAMPL5 compounds the energetically lowest structures of the neutral species were manually protonated and deprotonated at chemically plausible sites and reoptimized to relax the structure into the new local minimum. No further conformational searches were carried out for the charged species at this stage.

All 3D RISM calculations were conducted on cuboid grids with a grid spacing of 0.3 Å and a total grid size that is dependent on the size of the molecule in such a way that the atoms closest to the boxes edges were at least 35 Å away in the final 3D RISM calculation and at least 30 Å in all the preceding iterations of the EC-RISM cycle. One of the major differences between calculations conducted in this chapter and from chapter 5 onwards is the use of solute atomic charges generated from the wave function using CHelpG [84]. Beginning in chapter 5 and for the remainder of this work after that, instead the exact electrostatic potential generated from the wave function will be used after investigating their relative performance. Periodic boundary conditions were implemented via an Ewald summation scheme, where the potential is split into a short range real-space potential, that can be easily truncated, and a long range reciprocal-space potential [83]. The convergence criterion for the 3D RISM calculations was a maximum residual norm for the direct correlation functions of 10^{-6} between two iterations for the calculation to be considered converged. For cyclohexane, the model developed as described in the preceding chapter was used, while for water the model originally used in the SAMPL2 challenge without a PMV correction was used unchanged [43]. This model is derived from the SPC/E water model with the hydrogens modified to have a σ of 1.0 Å [144,145]. The PSE-2 closure was used for both solvents to make sure that all molecules of the SAMPL5 data set would converge while at the same time yielding accurate thermodynamic data.

In the EC-RISM calculations the HF/6-311+G(d,p) level of theory was used during the iterations until the change in the total Gibbs energy measured as the sum of electronic energy and excess chemical potential fell below 0.01 kcal mol^{-1}. After that a single iteration with the MP2/6-311+G(d,p) level of theory was performed to account for electron correlation effects. For all EC-RISM QM calculations in this chapter Gaussian 03 rev D.02 was used [146].

During the SAMPL5 challenge the PMV calculated from the total correlation function and the experimental isothermal compressibility were used (eq. (37)). This results in significantly differing calculated PMVs, but due to the scaling introduced by the PMV correction parameter c_V only minor changes in the corrected excess chemical potentials. Nevertheless, in the course of this work it was discovered that the PMVs calculated from the direct correlation function and using the RISM estimate for the compressibility (eq. (39)) should be used to achieve the most thermodynamically consistent results.

4.3 Results

4.3.1 Solvation free energies

In this work 481 of the molecules for which experimental Gibbs energies of hydration are available in the MNSOL database were used to train a PMV correction for water. This set contained 351 neutral compounds, 80 anions and water-anion clusters, and 50 cations and cation-water clusters. Additionally, 90 neutral molecules for which experimental Gibbs energies of solvation in cyclohexane were available were used to train the corrections for this solvent. For water a single model using three adjustable parameters, c_μ, c_V, and c_q was applied, because this model had shown good performance on the MNSOL dataset while it did not use many *ad hoc* parameters that might be the cause of overfitting [43]. Due to the lack of experience in working with cyclohexane as a solvent, two different models were derived from eq. (46) and tested here using the SAMPL5 challenge compounds as a benchmark set. The first one used three parameters c_μ, c_v, and a linear offset c_d and will be referred to as “3-par-I(5)” in the following, while the second model used only the two latter parameters and will be referred to as “2-par-I(5)”. The corrected excess chemical potential is then

$$\mu_\mathrm{corr}^\mathrm{ex} = c_\mu \mu^\mathrm{ex} + c_v V^m + c_o \tag{75}$$

for the 3-par-I(5) model and

$$\mu_\mathrm{corr}^\mathrm{ex} = \mu^\mathrm{ex} + c_v V^m + c_o \tag{76}$$

for the 2-par-I(5) model. The resulting Gibbs energies of solvation in water and cyclohexane in comparison with the experimental values are depicted in Fig. 3 while the optimized parameters and statistical metrics of the different models are shown in Table 3.

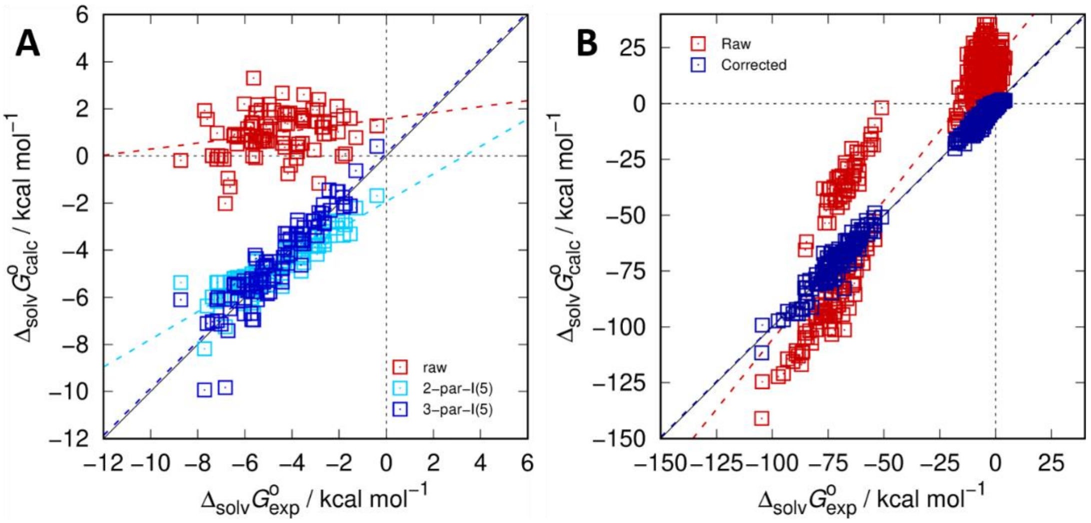


Fig. 3: Calculated versus experimental solvation free energies for the MNSOL dataset [14] using EC-RISM at the MP2/6-311+G(d,p) level of theory before (red) and after (blue) correcting the raw data using the corrections based on the partial molar volumes. Results are shown for cyclohexane (A) using the 3-par-I(5) model (blue) and the 2-par-I(5) model (light blue) and water (B). Dashed lines indicate descriptive regression results. Original uncorrected and corrected EC-RISM/3D RISM data are provided in OR_02, optimized solution and gas phase structures are collected in OR_03. Figure adapted from [1].

Table 3: Regression parameters of optimized EC-RISM-based Gibbs energy of solvation models (c_μ is unitless, c_V / kcal mol^{-1} Å^{-3}, c_q / kcal mol^{-1} e^{-1}, c_d / kcal mol^{-1}) along with statistical metrics (root-mean-square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*’, intercept *b*’ / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression). c_V was calculated using the PMVs calculated via the total correlation route [76], and the experimental isothermal compressibilities of 0.450183·10^{-9} Pa^{-1} for water and 1.1197·10^{-9} Pa^{-1} for cyclohexane [125]. Table adapted from [1].

Solvent	RMSE	MAE	MSE	*m*'	*b*’	R^2	c_μ	c_V	c_q	c_d
Water										
Uncorrected	20.84	18.83	-13.23	1.24	18.74	0.88	-	-	-	-
All	2.43	1.69	0.35	0.99	-0.52	0.99	0.97	-0.17	-17.26	-
Neutrals	1.52	1.17	-0.52	0.98	-0.62	0.89	-	-	-	-

Anions	4.48	3.78	-0.21	1.11	7.92	0.88	-	-	-	-
Cations	2.91	2.09	0.68	0.87	-7.95	0.85	-	-	-	-
Cyclohexane										
Uncorrected	5.90	5.65	-5.65	0.13	1.58	0.06	-	-	-	-
2-par-I(5)	0.88	0.70	0.00	0.59	-1.94	0.77	-	-0.11	-	-1.66
3-par-I(5)	0.76	0.48	0.00	0.84	-0.73	0.84	1.84	-0.15	-	-1.05

Here and throughout this work the descriptive regression metrics must not be mistaken for the parameters *m* and *b* resulting from the fit of the p*K*a models. A slope close to 1 and an offset close to 0 indicate a good model, provided that the R^2 is close to 1 as well. As can be seen in Fig. 3, after training of the PMV corrections the corrected Gibbs energies of solvation for both solvents are significantly closer to the experimental values than before.

The exact results of the PMV correction are summarized in Table 3. Using just three parameters for water the calculated hydration free energies exhibit a total RMSE of 2.43 kcal·mol^{-1}. This error is mostly determined by the performance of the ions, because while they are fewer in number, their RMSE is significantly larger: looking only at the neutral compounds the RMSE shrinks to 1.52 kcal·mol^{-1} while the errors for cations (2.91 kcal·mol^{-1}) and especially for anions (4.48 kcal·mol^{-1}) are significantly larger.

The results for the prediction of Gibbs energies of solvation in cyclohexane are slightly better for both models that were trained: the total RMSE is only 0.76 kcal·mol^{-1} for the three-parameter model and 0.88 kcal·mol^{-1} for the two-parameter model. This, in addition to the fact that every calculated statistical metric is improved in the training set data could be expected, because an additional parameter can only improve the residual (otherwise it should be set to 1 for the parameter directly scaling the excess chemical potential and 0 for all other parameters during training). Interestingly, while the parameter directly scaling the uncorrected 3D RISM excess chemical potential for the water model is close to one, implying only minor corrections to the raw value, using this as an adjustable parameter for cyclohexane leads to a value of 1.84 which leads to an almost doubled excess chemical potential before the correction takes place. This implies that either some of the physical assumptions made for the PMV correction are wrong or incomplete, there are some deficiencies in the cyclohexane model used, or the experimental data is not reliable. However, this large parameter value appears to be necessary to improve the slope of the descriptive regression which is far from unity for the 2-par-I(5) model. To investigate this both models were applied to the SAMPL5 challenge compounds, as this is a data set that contains none of the MNSOL compounds and is thus ideal to use as a true test set for the models.

4.3.2 p*K*a prediction

For the prediction of acidity constants, 103 pairs of acids and conjugated bases (or vice versa) from the experimental dataset by Klicić et al. [143] were used to train a model that is used to calculate pK_a values without using the Gibbs free energy of the proton in solution. Here, two different models were tested as well. While in this case the loss function for both models was the same, for the first model all "acids" and "bases" defined as pairs consisting of a neutral compound and a deprotonated form or a neutral compound and a protonated form, respectively were adjusted with a single set of parameters. In the second model three sets of parameters were generated, one for all acids, one for secondary and tertiary amines, and one for all other basic compounds as defined above. When using this split there are 51 acidic compounds, 17 secondary and tertiary amines, and 35 other basic compounds. Similar splits have been used in other works because it has shown to improve the predictive power of pK_a models even though the Gibbs free energy of the proton should be the same in all cases [147]. The resulting acidity constants in comparison with the experimental values are depicted in Fig. 4 while the optimized parameters and statistical metrics of the different models are summarized in Table 4.

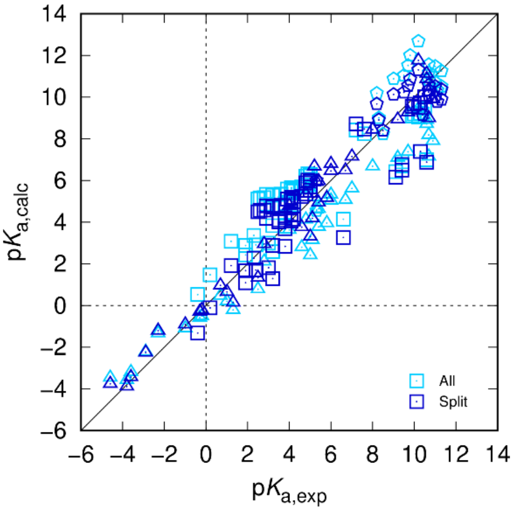


Fig. 4: pK_a values calculated using EC RISM at the MP2/6-311+G(d,p) level of theory compared with experimental values from the Klicić dataset. Data generated using individual 2-parameter regression models for acids, secondary and tertiary amines, and other bases is depicted in dark blue while data generated using a single model for all classes is depicted in light blue. Acids are depicted using squares, secondary and tertiary amines with pentagons and all other bases with triangles. Original data are provided as OR_02, optimized structures are collected in OR_03. Figure adapted from [1].

Table 4: Parameters *m* and *b* of optimized EC-RISM-based pK_a models and statistical metrics (root-mean-square error RMSE, mean absolute error MAE, mean signed error MSE, and coefficient of determination R^2 from descriptive regression). Table adapted from [1].

pK_a model	RMSE	MAE	R^2	*m*	*b*

All	1.52	1.29	0.85	0.74	-149.6
Acids	1.41	1.10	0.76	0.94	-191.7
Sec/tert amines	0.94	0.85	0.60	0.48	-94.6
Other bases	0.79	0.64	0.97	0.89	-149.6

As expected, the data in Table 4 shows that at least for the training set the split models perform significantly better than the unified model with a total RMSE of 1.15 and 1.52 pK units, respectively. The obtained parameters also appear to validate the choice of splitting the bases into secondary and tertiary amines, and all other kinds of bases. The obtained slope parameter for the amines is 0.48 and the offset only -94.6, while ideally these values should be 1, and a value corresponding to the Gibbs free energy of solvation of the proton, respectively. This is almost the case for the acids where the training yields a slope of 0.94 and an offset of -191.7 and the non-amine bases. Multiplying the unitless offset for this model with $RT\ln(10)$ yields an energy of -261.53 kcal mol^{-1}. For comparison, the Gibbs free energy of solvation of the proton has been experimentally determined as -265.89 kcal mol^{-1} [103]. Even visually the data in Fig. 4 suggests that the unified model has more outliers for all compound classes. Regardless, in the application of the model to the SAMPL5 dataset the unified model was chosen, because for the training set the compounds contain only a single ionizable group that can be clearly determined to belong to one of the three classes. This is not generally the case for the compounds contained in the SAMPL challenges, which often contain multiple ionizable groups that may be in different compound classes, making it difficult to properly assign which model should be used for which compound.

4.3.3 Prediction of distribution coefficients

For the SAMPL5 challenge not only was the choice made to only evaluate the model using a single set of parameters for all compound classes, but also to only use the pK_a value that has the greatest influence when applied with eq. (61) or (62), respectively in the case of multiple pK_a values for the same molecule. This is inaccurate only if an acidic and a basic pKa value of the same molecule are equally distant to 7.4 which is unlikely.

The PMV corrections and the pK_a model described above where applied to the batch 0 and batch 1 of the SAMPL5 dataset to predict the log $D_{7.4}$. Consequently, there are six different combinations of models to be explored by combining one water model, two cyclohexane models (2-par-I(5) and 3-par(5)), and three pK_a models (no pK_a, i.e. a log P model, the MoKa pK_a values, and the EC-RISM pK_a values). In the original SAMPL5 paper the batch 1 compounds SAMPL5_010, SAMPL5_011, SAMPL5_026, and SAMPL5_060 were erroneously treated as bases for the calculation of the distribution coefficient but are correctly treated as acids in this work. The calculated partition or distribution coefficients in comparison with the experimental values for the distribution coefficients are depicted in Fig. 5 while the optimized parameters and statistical metrics of the different models are shown in Table 5.

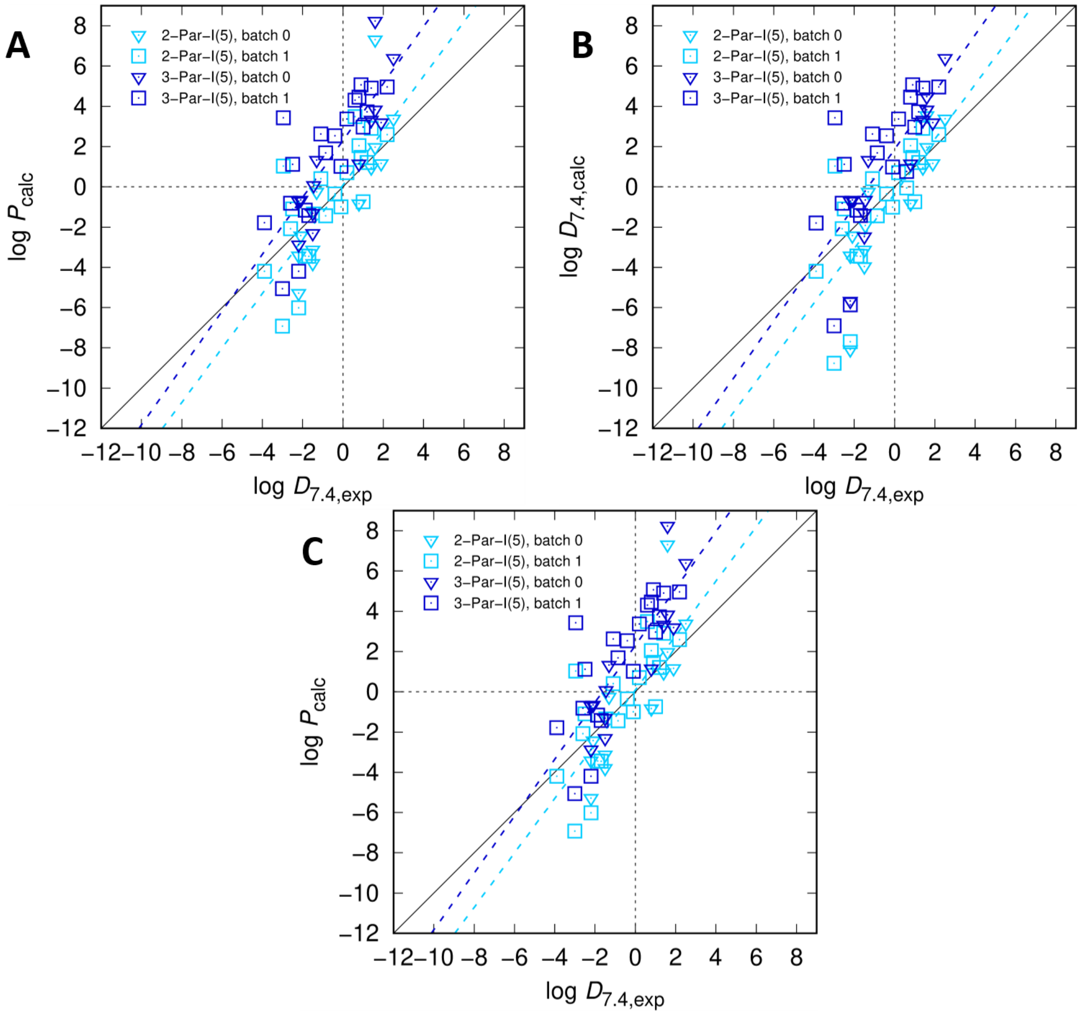


Fig. 5: Calculated partition coefficients using EC-RISM at the MP2/6-311+G(d,p) level of theory considering only neutral species (A) and distribution coefficients taking into account protonation equilibria using pK_a values derived from EC-RISM (B) or Moka (C). Results for compounds belonging to batch 0 are depicted in light blue while those for batch 1 are depicted in dark blue. Cyclohexane models using two parameters are depicted in light blue while those using three parameters are depicted in dark blue. Descriptive regression lines are depicted for batch 0 (light blue) and batch 1 (dark blue). Compounds where the most relevant ionized state is protonated are depicted as triangles (bases) and those where the most relevant ionized state is deprotonated as squares (acids). Original data are provided as part of OR_02, optimized structures are collected in OR_03. Figure adapted from [1].

Table 5: Statistical metrics (root-mean-square error RMSE, mean absolute error MAE, mean signed error MSE, and slope m, intercept b, and coefficient of determination R^2 from descriptive regression) for both partition and distribution coefficient predictions for the SAMPL5 compounds of batches 0 and 1. Table adapted from [1].

pK_a	cyclohexane	RMSE	MAE	MSE	R^2	m'	b'
None	2-par-I(5)	1.99	1.48	-0.09	0.61	1.35	0.09
	3-par(5)	2.86	2.08	2.08	0.65	1.41	2.30
EC-RISM	2-par-I(5)[a]	2.25	-0.86	1.63	0.71	1.60	-0.54
	3-par(5)[b]	2.59	1.31	2.29	0.70	1.66	1.67
Moka	2-par-I(5)	4.61	4.09	-3.12	0.23	1.12	-5.23

	3-par(5)	6.42	5.87	-5.30	0.28	1.19	-3.03

[a-b]Corrected results for SAMPL5 setup, original values [1] for RMSE, MSE, R^2, *m*', *b*':
[a]2.15, -0.53, 0.59, 1.36, -0.34;
[b]2.76, 1.64, 0.59, 1.42, 1.87.

Looking at the differences between the predictions shown in Fig. 5 the first observation that can be made is that the use of MoKa pK_a values (C) significantly shifts almost every predicted distribution coefficient to lower values compared to the log *P* predictions (A). This behavior can also be observed for the distribution coefficients calculated using acidity constants generated with EC-RISM (B), but there only for a few compounds, e.g. SAMPL5_015 and SAMPL5_072 (see Table 6). Interestingly, the use of acidity constants of either kind does not consistently improve the predictive power of the model. While for the 3-par(5) cyclohexane model a small improvement in the RMSE can be observed when using EC-RISM pK_a values, for the 2-par-I(5) model the opposite behavior is found.

The different cyclohexane models on the other hand exhibit a more conclusive picture. Ignoring the clearly inferior distribution coefficients calculated using MoKa pK_a values the 2-par-I(5) model has a significantly better performance predicting the SAMPL5 distribution coefficients than the 3-par(5) model with RMSEs of 1.99 and 2.86 for the log *P* models, and 2.15 and 2.76 for the log $D_{7.4}$ models, respectively. This is the opposite result compared to the one obtained in the training set and warrants closer examination.

The most important thing to note is that while the slope for results obtained using either cyclohexane model is almost unchanged, the offset of the regression is shifted towards lower values by approximately 2.2 pK units for the 2-par-I(5) models. In other words, the 3-par(5) models tend to overestimate the fraction of the compounds present in the organic phase. This can also be recognized in the figure where the central cluster of data points lies on the diagonal for the 2-par-I(5) models, but significantly above it for the 3-par(5) models, leading to a higher RMSE and MSE overall.

Another counter-intuitive result is the fact that application of the pK_a model to calculate distribution coefficients leads to worse results than achieved by the calculated partition coefficients. This can be seen in Table 6, where a closer look at the results reveals that while for some compounds that were previously estimated as too lipophilic but have an ionizable group the predicted log $D_{7.4}$ is better than the predicted log *P*. An example for this is the compound SAMPL5_070, where the log *P* model fails completely with a deviation of 6.65, the structure of which is depicted in Fig. 6.

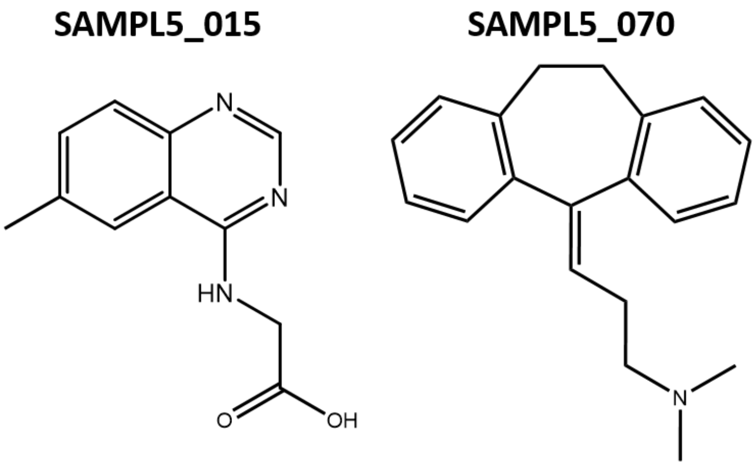


Fig. 6: Chemical structure of the compound SAMPL5_070 and SAMPL5_015 in their neutral form. Only the most abundant tautomer is depicted if multiple tautomeric states are possible.

Such a high predicted partition coefficient is actually plausible, because the molecule is almost completely apolar, but the dimethylamine is predicted to be almost completely in its protonated form at pH 7.4, which leads to a predicted log $D_{7.4}$ that is nearly four orders of magnitude lower, leading to a deviation of only 2.88. This is still quite large, but much better than the originally predicted neutral state partition coefficient.

However, the exact opposite trend can be observed for some compounds that were already predicted to be too hydrophilic, and these molecules are then predicted even worse than before. One example of this is the compound SAMPL5_015 that is also shown in Fig. 6 and is predicted quite well, with only a deviation of -0.67 pK units, but the carboxylic acid is predicted to be mostly deprotonated at pH 7.4 with an acid dissociation constant of 4.60. A full list of all calculated partition and distribution coefficients for the batch 0 and batch 1 SAMPL5 compounds as well as their experimental distribution coefficient is provided in Table 6.

Table 6: Individual experimental and computational data for distribution coefficients underlying the statistics shown in Fig. 5 and Table 5. Asterisks denote compounds with acidic pK_a values where the log D is evaluated according to eq. (62), all other compounds had basic pK_a values and the log D is evaluated according to eq. (61).

SAMPL5 ID	log $D_{7.4,\mathrm{exp}}$	log P 2-par-I(5)	log P 3-par(5)	p$K_{\mathrm{a,MoKa}}$	log $D_{7.4,\mathrm{MoKa}}$ 2-par-I(5)	log $D_{7.4,\mathrm{MoKa}}$ 3-par(5)	p$K_{\mathrm{a,ECR}}$	log $D_{7.4,\mathrm{ECR}}$ 2-par-I(5)	log $D_{7.4,\mathrm{ECR}}$ 3-par(5)
Batch 0									
003	1.90	1.17	3.19	15.42	-6.85	-4.83	-3.56	1.17	3.19
015	-2.20	-5.28	-2.87	15.31	-13.19	-10.78	4.60*	-8.08	-5.67
017	2.50	3.39	6.39	15.09	-4.30	-1.30	-0.26	3.39	6.39
020	1.60	1.98	3.83	13.32	-3.94	-2.09	0.67	1.98	3.83
037	-1.50	-3.79	-2.31	4.38	-3.79	-2.31	7.05	-3.95	-2.47
045	-2.10	-2.42	-0.64	12.04	-7.06	-5.28	0.91	-2.42	-0.64
055	-1.50	-3.13	-1.31	14.11	-9.84	-8.02	-1.24	-3.13	-1.31

058	0.80	-0.83	1.16	16.00	-9.43	-7.44	-5.84	-0.83	1.16
059	-1.30	-0.25	1.32	12.73	-5.58	-4.01	0.52	-0.25	1.32
061	-1.45	-1.19	0.08	15.00	-8.79	-7.52	8.04	-1.91	-0.65
068	1.40	0.95	3.33	15.00	-6.65	-4.27	3.05	0.95	3.33
070	1.60	7.32	8.25	5.31	7.32	8.24	11.17	3.56	4.48
080	-2.20	-3.42	-0.71	12.02	-8.04	-5.33	-0.95	-3.42	-0.71
Batch 1									
004	2.20	2.60	4.96	7.08	2.43	4.79	0.05	2.60	4.96
005	-0.86	-1.44	1.68	8.71	-2.77	0.35	3.84	-1.44	1.68
007	1.40	2.91	4.90	15.14	-4.83	-2.84	3.63	2.91	4.90
010[a]	-1.70	-3.45	-1.43	14.47	-10.52	-8.50	4.98*	-5.88	-1.43
011[b]	-2.96	1.03	3.43	14.48	-6.05	-3.65	4.71*	-1.67	3.43
021	1.20	1.22	3.72	14.34	-5.72	-3.22	3.06	1.22	3.72
026[c]	-2.60	-2.08	-0.82	4.47*	-2.08	-0.82	4.46*	-5.02	-0.82
027	-1.87	-3.44	-1.16	14.53	-10.57	-8.29	2.26	-3.44	-1.16
042	-1.10	0.40	2.63	16.00	-8.20	-5.97	-2.89	0.40	2.63
044	1.00	-0.74	2.97	11.58	-4.92	-1.21	0.36	-0.74	2.97
046	0.20	0.70	3.38	16.87	-8.77	-6.09	-1.55	0.70	3.38
047	-0.40	-0.35	2.53	15.00	-7.95	-5.07	-8.41	-0.35	2.53
048	0.90	1.47	5.07	15.00	-6.13	-2.53	0.81	1.47	5.07
056	-2.50	-1.10	1.12	16.00	-9.70	-7.48	-2.58	-1.10	1.12
060[d]	-3.90	-4.19	-1.79	3.31*	-8.28	-5.88	4.74*	-6.86	-1.79
063	-3.00	-6.39	-5.06	8.36	-7.93	-6.06	9.24	-8.77	-6.90
071	-0.10	-0.99	1.02	10.05	-3.64	-1.63	6.12	-1.02	0.99
072	0.60	3.49	4.30	5.45	3.49	4.30	10.94	-0.05	0.76
081	-2.20	-6.02	-4.20	4.81	-6.03	-4.20	9.05	-7.69	-5.86
090	0.80	2.04	4.46	13.33	-3.89	-1.47	2.53	2.04	4.46

[a-d]Corrected results for SAMPL5 setup, original data [1] for log $D_{7.4,\mathrm{MoKa}}$(2-par-I(5), 3-par) and log $D_{7.4,\mathrm{EC\text{-}RISM}}$(2-par-I(5), 3-par):
[a]-10.52, -8.50, -3.45, -1.43;
[b]-6.05, -3.65, 1.03, 3.43;
[c]-2.08, -0.82, -2.08, -0.82;
[d]-8.28, -5.88, -4.19, -1.79.

Some possible explanations for these deviations exist: For one, the system investigated experimentally may differ from the system with which the Gibbs free energies of solvation were originally measured. For example, the experimental conditions for the SAMPL5 challenge included cosolvents such as 1% DMSO and 0.5% acetonitrile, phosphate buffer for the aqueous phase and octanol to dilute the injection volume [52]. None of these additional substances were represented in the water or cyclohexane models. Another reason may be that overfitting to the Gibbs energies of solvation by doubling the 3D RISM excess chemical potential when applying the correction leads to overestimated excess chemical potentials corresponding to overestimated partitioning into the organic phase. However, due to the lack of further, difficult to obtain experimental data the error in the partitioning cannot be separated into its components. For this reason, it is at this point impossible to know if the error in the calculated distribution coefficients is caused by errors in the Gibbs energies in one of the two or in both solvents. The lack of experimental acidity constants for the individual compounds also makes it difficult to ascertain to what degree the significant discrepancies between partition and distribution coefficient would be found in experimental measurements. Further discussion and analysis of a reevaluation of the entire SAMPL5 set of compounds will be conducted in Chapter 6.4.

5 SAMPL6.1: PREDICTION OF ACIDITY CONSTANTS IN AQUEOUS SOLUTION

5.1 Introduction

The task set by the authors in the SAMPL5 challenge was quickly recognized as being perhaps slightly too difficult compared to the “simple” prediction of hydration free energies in the earlier SAMPL challenges. Especially the fact that errors cannot be properly assigned to the modeling of either phase or the pK_a model makes it difficult to analyze and improve the models used. For example, if a model performs well in predicting the energy in the water phase but significantly worse for the octanol phase the results are indistinguishable from those of a model for which the opposite is true.

For this reason, the SAMPL6 challenge was created to tackle the same task, the distribution of compounds between two immiscible phases, but split into two parts: the first task was to predict the acidity constants of 24 compounds in aqueous solution, while at a later date the prediction of the octanol-water partition coefficient of the neutral form of the molecules would form an independent challenge. Furthermore, even the first task itself was split into three parts: the prediction of microscopic pK_a values, that is, the acid dissociation constant between two specific microstates of a given compound, the prediction of microstate populations as a function of the pH, and finally the prediction of macroscopic, experimentally measurable pK_a values. Of these only the last task had full experimental results available, while the microstate populations were only determined for a few, selected compounds. Unlike in the previous challenge, here, a comprehensive set of tautomeric states in all potentially relevant protonation states (sometimes ranging from -3 to +4) was issued by the challenge authors.

This task is perfectly suited for EC-RISM for a number of reasons: for one, all the required predictions can be made by simply calculating the Gibbs energies of all relevant tautomers in each protonation state under investigation. Furthermore, the use of quantum-chemical methods should make the calculations more accurate, especially for the highly charged compounds, as polarization effects become more important. Three different quantum-chemical levels of theory were investigated to learn about their relative performance for predicting physicochemical parameters. Finally, compared to the earlier SAMPL5 challenge there were some improvements that could be tested for the first time in this blind prediction challenge. During the SAMPL5 challenge the electrostatic potential used in the EC-RISM calculations was derived from atomic partial charges, while here the exact electrostatic potential was derived from the wave function of the solute, which should be more accurate. In addition to this, over the course of the challenge P. Kibies *et al.* developed a truncation scheme for the exact electrostatic potential, because of convergence issues for some compounds when using it during EC-RISM calculations. The problem is caused by the exact electrostatic potential not being sufficiently close to the point-charge based potential at the box edges, especially for some charged compounds. P. Kibies' scheme scales the exact electrostatic potential with a cubic switching function [87]. For comparison's sake a simple pK_a model taking the PCM energies at the B3LYP/6-311+G(d,p) level of theory as input was also developed and submitted in the SAMPL6 challenge to investigate if it is possible to gain access to reasonable acid dissociation constants at such little computational cost.

5.2 Computational details

In the interest of brevity, in the following chapters only those computational details that differ from the setup used in the SAMPL5 challenge will be mentioned. In this instance only the globally minimal PCM structures were taken to calculate the Gibbs energies of solvation and the acidity constants for the MNSOL training set [17]. For the SAMPL6 challenge set a more efficient way of generating the conformers was used. This may seem counterintuitive considering that only 24 molecules had to be investigated, however, considering that the sum of all microstates published by the authors made this number swell to 413 distinct states, for each of which the conformational ensemble had to be generated, this course of action was necessary. Thus, to generate the conformations Dr. Stefan Güssregen from our collaboration partner the Sanofi-Aventis Deutschland GmbH used Maestro 11.2 and Macromodel 11.6 from the 2017-2 release of the Schrödinger software suite using OPLS3 with an implicit water model and the mixed torsional/low-mode conformational search algorithm [148]. The search was carried out with 100 steps for each rotatable bond up to a maximum of 1000 steps. An energy cutoff of 5 kcal·mol^{-1} and an RMSD cutoff of 1.5 Å were used to eliminate high-energy and redundant conformations from the search results. In one compound, SM22, iodine was present. Since this element cannot be treated with all basis sets used in this work it was replaced by the next smaller halogen, bromine, for every microstate. For the challenge submission only the minimum conformation supplied by Macromodel was optimized quantum-chemically, first using PBE/6-311+G(d,p) for pre-optimization followed by an optimization at the same level of theory used in the SAMPL5 challenge. After the submission all conformations were optimized this way and the lowest two conformations as measured by their PCM energies were calculated with EC-RISM. During the EC-RISM calculations three different levels of theory were used: the SAMPL5 level of theory MP2/6-311+G(d,p), the slightly more expensive level of theory MP2/cc-pVTZ, and the inexpensive DFT functional B3LYP/6-311+G(d,p). All calculations, including the EC-RISM calculations, were conducted using Gaussian 09 rev E.01 [149].

Some of the microstates contained a protonated oxygen atom with a positive atomic charge and bonded to three different atoms. These atoms have no corresponding GAFF parameters and microstates containing them were discarded. Judging from their relative PCM energies compared to other microstates of the same protonation state this was not expected to have an effect on the final energy.

The 3D RISM calculations were conducted on size-adapted grids with the same 0.3 Å spacing as in the SAMPL5 challenge for the submission phase, however this time using distances of only 12.5 Å between the molecule and the box edges during the HF iterations and 15 Å in the final MP2 iteration. For the calculations after the initial submission cubic boxes with a fixed grid size of 128^3 points were used for all molecules other than SM23, where due to the large size of the molecule a grid of 140^3 points was used.

5.3 Results

5.3.1 Solvation free energies

In this challenge a greater number of models was tested to investigate if the choice of the MP2/6-311+G(d,p) level of theory in combination with a 3-par PMV correction for water and a 2-par correction for the pK_a were sensible. In addition, the new treatment of the electrostatics further increases the number of models under consideration. For the SAMPL6 challenge the following PMV corrections were trained:

1. Three levels of theory: MP2/cc-pVTZ, MP2/6-311+G(d,p), B3LYP/6-311+G(d,p). The 6-311+G(d,p) basis set will be denoted as “6-311” in the model designations
2. Three treatments of the electrostatics: point charges (“q”), full electrostatics without periodicity correction (“φ”), and full electrostatics with periodicity correction (“φ_{opt}”)
3. Two PMV corrections: including direct scaling of the 3D RISM excess chemical potential (3-par) and no scaling of the 3D RISM excess chemical potential (2-par). In this case neither model contains an intercept parameter, but instead both contain a parameter accounting for the Galvani potential when investigating charged species (see chapter 2.2.1).

Compared to the corrections used for cyclohexane (see equations (75) and (76)), in this case the corrected excess chemical potentials for the two corrections are

$$\mu_{\mathrm{corr}}^{\mathrm{ex}} = c_{\mu}\mu^{\mathrm{ex}} + c_{v}V^{m} + c_{q}q \tag{77}$$

for the 3-par model and

$$\mu_{\mathrm{corr}}^{\mathrm{ex}} = \mu^{\mathrm{ex}} + c_{v}V^{m} + c_{q}q \tag{78}$$

for the 2-par model. This gives a total of 18 possible PMV corrections under consideration of which only a few selected examples can be fully discussed in the main body of this work, because the different pK_{a} models further expand the number of potential models.

5.3.1.1 MP2-based models

First the Gibbs energies of hydration calculated with the MP2 level of theory and the 6-311+G(d,p) and cc-pVTZ basis sets will be discussed in full, as these were the models originally submitted in the SAMPL6 challenge [2]. The calculated Gibbs energies of solvation using point charges and the full electrostatic potential with periodicity correction are shown in Fig. 7 and the full sets of parameters and statistical metrics for this level of theory in Table 7.

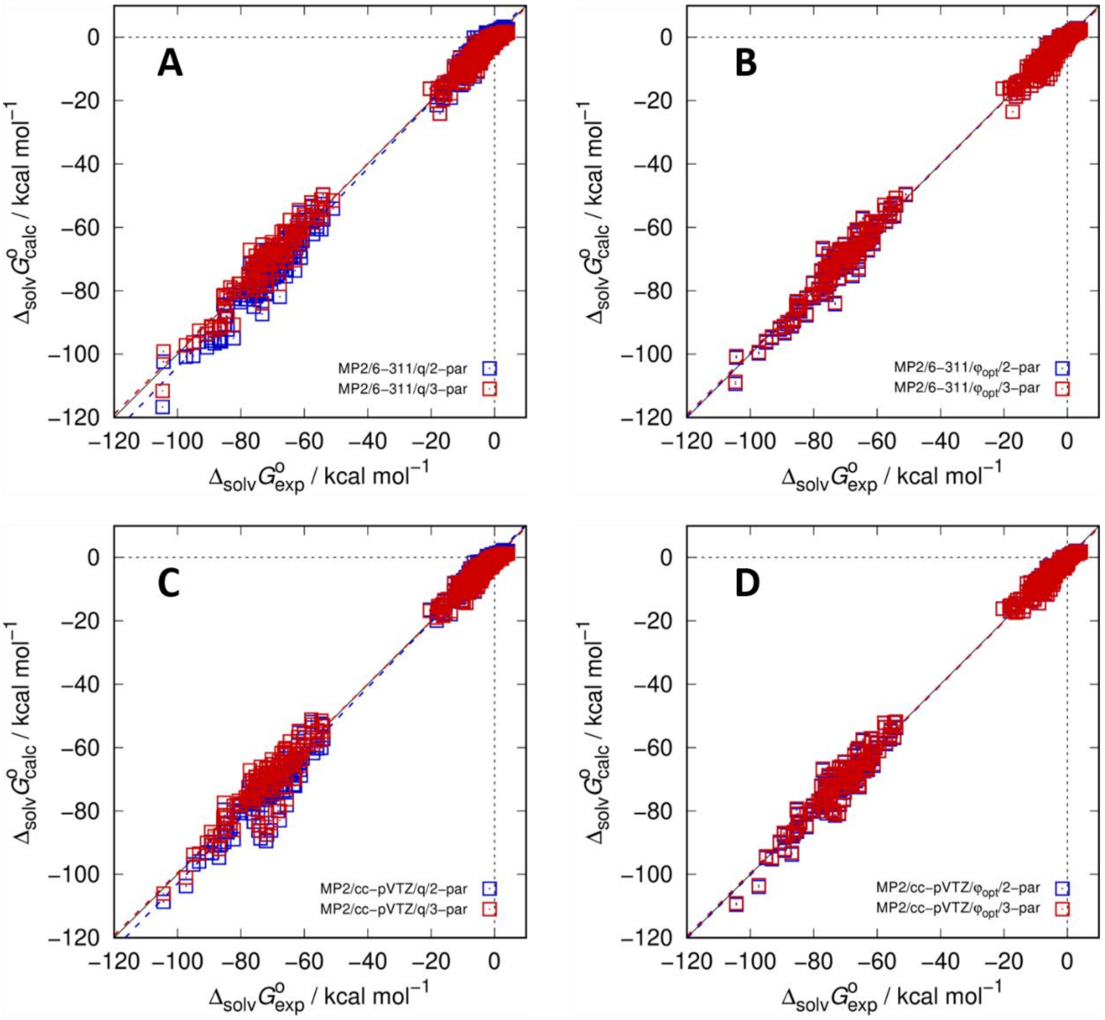


Fig. 7: Gibbs energies of solvation calculated using EC-RISM at the MP2 level of theory using the 6-311+G(d,p) (A,B) and the cc-pVTZ (C,D) basis sets vs. the experimental results from the MNSOL database using point charges (A,C) and the periodicity-corrected exact electrostatic potential (B,D). Data generated using the 2-par correction is shown using blue squares and data generated using the 3-par correction as red squares. Original data are provided as part of OR_04, optimized structures are collected in OR_05. The latter are the same for all models investigated in this chapter. Figure adapted from [2].

Table 7: Parameters of optimized EC-RISM-based solvation free energy models for water (c_μ, c_V / kcal mol^{-1} Å^{-3}, c_q / kcal mol^{-1} e^{-1}) along with statistical metrics (root mean square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*’, intercept *b*’ / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression). This data was generated using the experimental isothermal water compressibility of 0.450183·10^9 Pa^{-1}. Table adapted from [2].

EC-RISM	RMSE	MAE	MSE	*m*'	*b*’	R^2	c_μ	c_V	c_q
MP2/6-311/q/2-par	2.99	2.01	-0.56	1.04	0.42	0.99	-	-0.1608	-20.5422
Neutral	1.77	1.31	-0.20						
Anions	5.27	3.92	2.23						
Cations	4.66	4.16	3.61						
MP2/6-311/q/3-par	2.32	1.65	-0.09	0.99	-0.34	0.99	0.9538	-0.1623	-17.8117
Neutral	1.53	1.16	0.31						
Anions	4.17	3.62	-0.68						
Cations	2.93	2.13	-0.42						
MP2/6-311/φ/2-par	2.20	1.53	-0.29	1.00	-0.29	0.99		-0.1630	-16.0322
Neutral	1.63	1.17	-0.30						
Anions	3.50	2.76	-0.19						
Cations	3.04	2.19	-0.31						
MP2/6-311/φ/3-par	2.16	1.50	-0.19	0.99	-0.44	0.99	0.9901	-0.1633	-15.5270

Neutral	1.60	1.17	-0.40						
Anions	3.42	2.72	0.35						
Cations	3.04	2.03	0.56						
MP2/6-311/φ_{opt}/2-par	2.04	1.43	-0.26	1.00	-0.35	1.00	-	-0.1633	-15.7284
Neutral	1.56	1.13	0.36						
Anions	3.07	2.46	-0.01						
Cations	2.98	2.09	-0.02						
MP2/6-311/φ_{opt}/3-par	2.03	1.42	-0.20	0.99	-0.44	1.00	0.9938	-0.1634	-15.4259
Neutral	1.55	1.13	-0.42						
Anions	3.03	2.44	-0.35						
Cations	3.00	2.01	-0.54						
MP2/cc-pVTZ/q/2-par	3.04	1.69	-0.44	1.03	0.26	0.99	-	-0.1663	-19.4448
Neutral	1.42	1.04	0.13						
Anions	6.21	4.36	-1.69						
Cations	4.03	2.19	-2.80						
MP2/cc-pVTZ/q/3-par	2.72	1.69	-0.11	0.99	-0.29	0.99	0.9652	-0.1669	-17.5025
Neutral	1.37	1.05	-0.24						
Anions	5.68	4.36	0.23						
Cations	2.87	2.19	0.38						
MP2/cc-pVTZ/φ/2-par	2.32	1.47	-0.27	1.00	-0.26	0.99	-	-0.1679	-15.9468
Neutral	1.33	0.97	-0.26						
Anions	4.51	3.39	-0.26						
Cations	2.94	2.19	-0.41						
MP2/cc-pVTZ/φ/3-par	2.30	1.47	-0.27	0.99	-0.37	0.99	0.9937	-0.1677	-15.5604
Neutral	1.34	0.97	-0.26						
Anions	4.46	3.39	-0.26						
Cations	2.90	2.19	-0.41						
MP2/cc-pVTZ/φ_{opt}/2-par	2.20	1.41	-0.26	1.00	-0.28	0.99	-	-0.1677	-15.8696
Neutral	1.30	0.94	-0.28						
Anions	4.22	3.22	-0.17						
Cations	2.90	2.13	-0.28						
MP2/cc-pVTZ/φ_{opt}/3-par	2.19	1.41	-0.26	0.99	-0.37	0.99	0.9923	-0.1680	-15.5669
Neutral	1.31	0.94	-0.28						
Anions	4.17	3.22	-0.17						
Cations	2.89	2.13	-0.28						

Using no direct scaling of the excess chemical potential appears to make the results slightly worse when using point charges during the EC-RISM calculations. For the model MP2/6-311/q/2-par the RMSE reaches 2.99 kcal·mol^{-1} across all molecules while for the corresponding 3-par model achieves an RMSE of 2.32 kcal·mol^{-1}. The increased error is mostly caused by a worse prediction of anions and cations where the RMSE is 5.27 and 4.66 kcal·mol^{-1} instead of 4.17 and 2.93 kcal·mol^{-1}, respectively. This is not a surprise, because a less accurate model for the treatment of the electrostatics should have a larger influence on charged molecules, where the electrostatics contribute the most to the Gibbs energy of solvation. The model using point charges is also the only one where the trained parameter scales the calculated excess chemical potential by more than 1%. For the other models using the exact electrostatic potential the model parameter for scaling the excess chemical potential is close to unity and thus the differences in the errors and other statistical parameters between the 2-par and 3-par models is negligible. This is also the reason why only the 2-par models were discussed in the original paper [2].

Using the full QM electrostatics without any further correction already improves the results significantly. While for neutral species the RMSE decreases only to 1.63 kcal·mol^{-1}, the anions and cations are now predicted with an RMSE of 3.50 and 3.04 kcal·mol^{-1}, respectively, resulting in a total RMSE of only 2.20 kcal·mol^{-1}. Using the periodicity correction with the model MP2/6-311/φ_{opt}/2-par further decreases the RMSE to 2.04 kcal·mol^{-1}.

Looking at the results for the cc-pVTZ basis set some similarities and some differences are easy to make out. For this level of theory, the trends are the same for the point charge and electrostatic models and here, too, the parameter directly scaling the excess chemical potential is only significantly deviating from unity for the point charge model MP2/cc-pVTZ/q/3-par. Using the exact electrostatics improves the total RMSE from 3.04 to 2.32 kcal·mol^{-1} for the 2-par model and to 2.20 kcal·mol^{-1} with the MP2/cc-pVTZ/φ_{opt}/2-par model. But while for neutral and cationic species the result is slightly better than with the corresponding models using the 6-311+G(d,p) basis set, anions are predicted significantly worse here. Using the MP2/6-311/φ_{opt}/2-par model gives an RMSE of 3.07 kcal·mol^{-1} for the anions, which is almost the same as the RMSE for the cations of 2.98 kcal·mol^{-1}. For the cc-pVTZ model these values are 4.22 and 2.90 kcal·mol^{-1}, respectively, which shows a significant difference in the relative performance. Even at this point this should be expected to constitute a challenge when predicting the acidity constants of compounds where the deprotonated species is an anion.

Two important observations can be made just from this training set: The addition of a third free parameter does not appear to lead to improved results when using the exact electrostatic potential. It is thus possible to think of this parameter as a correction term for the approximation of using point charges to calculate the excess chemical potential. Furthermore, the usage of the full electrostatic potential mostly increases the accuracy for ions which means that for properties involving only neutral species, such as partition coefficients, the accuracy should be comparable while for properties involving charged species, such as acidity constants and distribution coefficients, the models using the full electrostatic potential should perform significantly better.

5.3.1.2 DFT-based models

While the following results were not published in the original paper, during the SAMPL6 challenge the DFT functional that is generally used to optimize the structures was also investigated for the quality of the EC-RISM energies. Furthermore, a simple PCM-based model that requires even fewer computational resources was developed and submitted as a model in the challenge. Compared to the MP2 calculations these DFT based energies can be calculated much faster because they scale much more favorably with the number of basis functions, making it possible to evaluate much larger or significantly more molecules at the same computational cost. The Gibbs energies of hydration calculated with the B3LYP/6-311+G(d,p) level of theory and using point charges and the full electrostatics with periodicity correction are exemplary depicted in Fig. 8 and the full sets of parameters and statistical metrics for this level of theory in Table 8. Fig. 8 and Table 8 also contain the data from the PCM model, where a single parameter correcting the Gibbs energy of hydration for ions, regardless of the sign of their total charge, was used. While this implicit solvation model does not have the same issues with respect to the PMV of the molecule as EC-RISM based models, the Gibbs energies of hydration of ionic compounds exhibited a strong deviation from the experimental values. A correction using only a single parameter multiplied by the absolute value of the net charge of the molecule was trained on the MNSOL database and applied to the PCM energies in all further calculations.

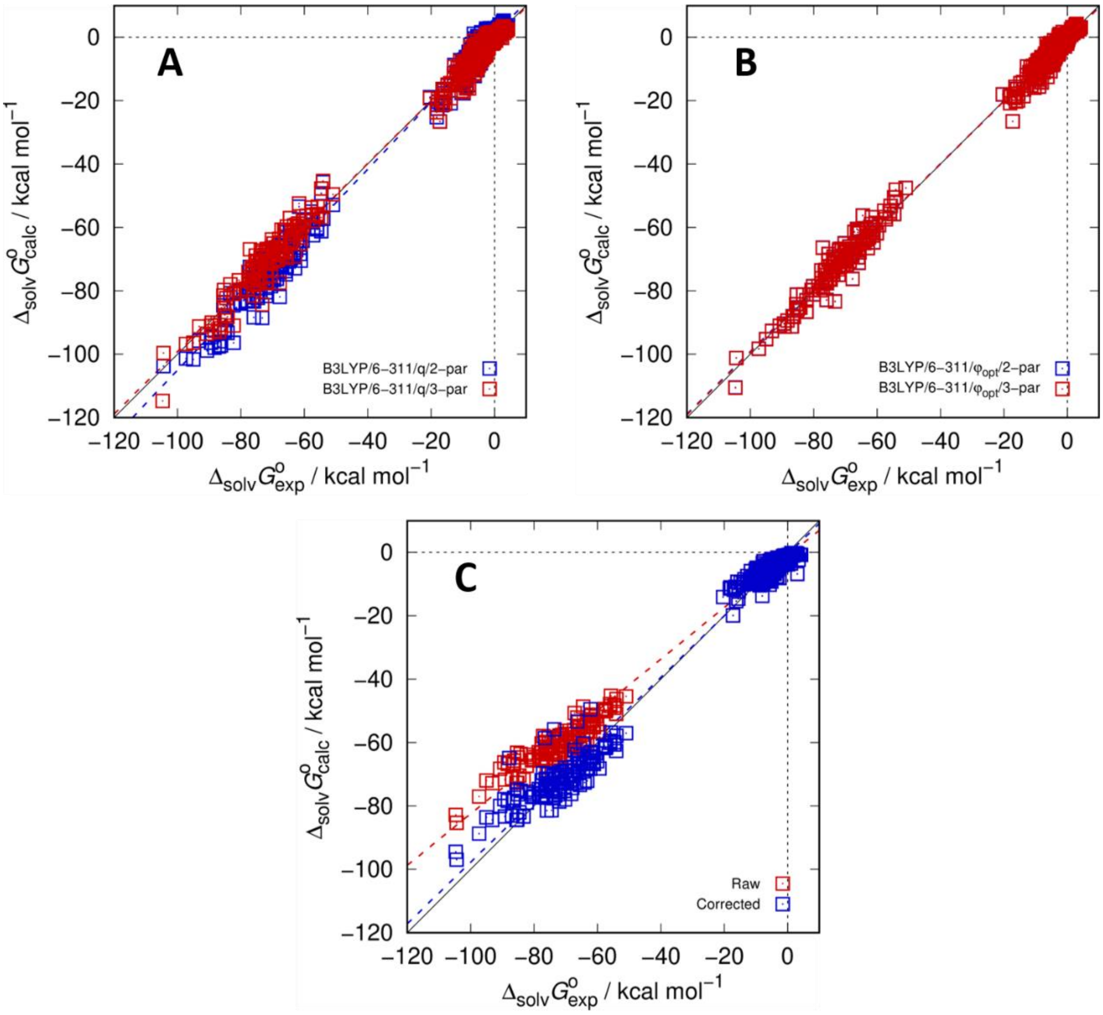


Fig. 8: Gibbs energies of solvation calculated using EC-RISM (A,B) and PCM (C) at the B3LYP/6-311+G(d,p) level of theory vs. the experimental results from the MNSOL database using point charges (A) and the periodicity-corrected exact electrostatic potential (B). For the EC-RISM-based models data generated using two free parameters is shown using blue squares and data generated using three free parameters as red squares. For the

PCM-based model uncorrected data is shown in red and corrected data in blue. Original data are provided as part of OR_04.

Table 8: Parameters of optimized EC-RISM- and PCM-based solvation free energy models for water (c_μ, c_V / kcal mol^{-1} Å^{-3}, c_q / kcal mol^{-1} e^{-1}) along with statistical metrics (root mean square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*', intercept *b*' / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression). For the PCM-based model only, the uncorrected data is shown as well. The EC-RISM-based data was generated using the experimental isothermal water compressibility of 0.450183·10^{-9} Pa^{-1}.

Model	RMSE	MAE	MSE	*m*'	*b*'	R^2	c_μ	c_V	c_q
EC-RISM									
B3LYP/6-311/q/2-par	3.56	2.54	-0.59	1.06	0.63	0.99	-	-0.1472	-22.7004
Neutral	2.47	1.88	0.30						
Anions	5.82	4.35	-2.54						
Cations	5.26	4.64	-4.20						
B3LYP/6-311/q/3-par	2.71	1.97	-0.02	0.99	-0.30	0.99	0.9426	-0.1482	-19.3426
Neutral	2.08	1.56	-0.33						
Anions	4.41	3.73	0.69						
Cations	3.17	2.28	1.14						
B3LYP/6-311/φ/2-par	2.32	1.65	-0.19	0.99	-0.33	0.99	-	-0.1541	-16.4628
Neutral	1.84	1.36	-0.35						
Anions	3.39	2.57	0.23						
Cations	3.20	2.28	0.38						
B3LYP/6-311/φ/3-par	2.31	1.63	-0.14	0.99	-0.40	0.99	0.9953	-0.1541	-16.2306
Neutral	1.83	1.35	-0.40						
Anions	3.37	2.54	0.48						
Cations	3.25	2.26	0.79						
B3LYP/6-311/φ_{opt}/2-par	2.19	1.54	-0.14	0.99	-0.39	0.99	-	-0.1546	-15.9949
Neutral	1.76	1.29	-0.41						
Anions	3.04	2.31	0.52						
Cations	3.23	2.25	0.86						
B3LYP/6-311/φ_{opt}/3-par	2.19	1.55	-0.14	0.99	-0.38	0.99	1.0009	-0.1546	-16.0396
Neutral	1.77	1.29	-0.40						
Anions	3.04	2.31	0.47						
Cations	3.21	2.25	0.78						
PCM (Uncorrected)	6.86	4.49	3.01	0.8134	-1.1402	0.99	-	-	-
PCM/6-311/2-par	3.57	2.53	0.06	0.9718	-0.5700	0.99	-	-	-11.6360
Neutral	3.02	2.06	0.08						
Anions	5.57	4.60	0.66						
Cations	3.22	2.73	-1.10						

The results are significantly worse when using point charges than for the equivalent models using MP2 energies, but the use of the full electrostatic potential closes most of this gap. This can be exemplified with the B3LYP/6-311/q/2-par model, which yields and RMSE of 3.56 kcal·mol^{-1} after training compared to the corresponding MP2 models that have an RMSE of only approximately 3.0 kcal·mol^{-1}. The B3LYP/6-311/φ_{opt}/2-par model on the other hand achieves an RMSE of merely 2.19 kcal·mol^{-1} which is well in line with the MP2 models' performances. Due to the bad performance for anions when using the MP2/cc-pVTZ level of theory the DFT approach in practice yields predictions with similar accuracy as these models, e.g., the MP2/cc-pVTZ/φ_{opt}/2-par has an almost identical RMSE of 2.20 kcal·mol^{-1}. And while for the MP2/cc-pVTZ model the performance varies significantly for cations and anions, this is not the case for the DFT models, making the use of DFT energies potentially advantageous for pK_a prediction, even though the performance of the MP2/6-311 model can still be expected to be superior.

The performance of the PCM model similar to the B3LYP/6-311/q/2-par EC-RISM model which is the worst-performing EC-RISM model investigated in this section, with total RMSEs of 3.56 and 3.57 kcal·mol^{-1}, respectively. As in the case of the MP2/cc-pVTZ models, here the significant difference in the performance for anions and cations is a concern when aiming to use a single set of parameters for the calculation of acidity constants, too.

Both the MP2- and the DFT-based corrections can now be applied to train pK_a models so their relative performance can be re-evaluated on this independent data set to investigate if the training set performance can be considered predictive.

5.3.2 pK_a model training

Unlike during the SAMPL5 challenge, here, the pK_a models were also investigated in more details. While the split of the training set into acids, secondary and tertiary amines, and other bases did not show any improvement in the predicted distribution coefficients and is not used here for that reason, there is in theory no reason for the slope parameter to be anything but “1”. However, as seen in this and other groups’ works [143,147,150], the parameter seems to be necessary to achieve predictions with chemical accuracy. Nevertheless, it is useful to know how much of an influence this additional parameter has on the results and how much this influence changes depending on the level of theory and the PMV correction used. In addition to the “2-par” and “3-par” for the PMV correction, these models are denoted “1-par” and “2-par”, respectively. Furthermore, during the training of the pK_a models one compound class, thiols, stood out as having significantly worse predicted pK_a values. The reason for this was later determined to be that the σ-Lennard Jones parameter of the negatively charged sulfur atom used in the GAFF parameter set was too small [151,152], but at the time of the challenge for each pK_a model two versions were trained: one with all molecules of the Klicić data set for training denoted “all”, and one excluding all thiol compounds from the regression denoted “nt”. Ultimately this means that for every PMV model four pK_a models were trained, but in the original paper only five of the models were discussed, namely the MP2/6-311/*/2-par/all models, where “*” denotes any of the possible options, as well as the MP2/cc-pVTZ/q/2-par/all and the MP2/cc-pVTZ/φ_{opt}/2-par/all models. This notation for all possible variations of a model designation will be used in the rest of this chapter as well. However, in this chapter the entire spectrum of different model combinations is shown, and the general trends resulting from each option will be discussed through some examples.

5.3.2.1 MP2-based models

First, the effect of the exact electrostatics, with and without periodicity correction will be compared with the point charge models using both the 6-311+G(d,p) and the cc-pVTZ basis sets that were discussed for the SAMPL6 challenge [2]. Furthermore, the influence of the two different PMV correction models, 2-par and 3-par, will be investigated. The calculated acidity constants of the training set for these points of discussion are shown in Fig. 9 and the parameters for the pK_a models and the resulting statistical metrics in Table 9.

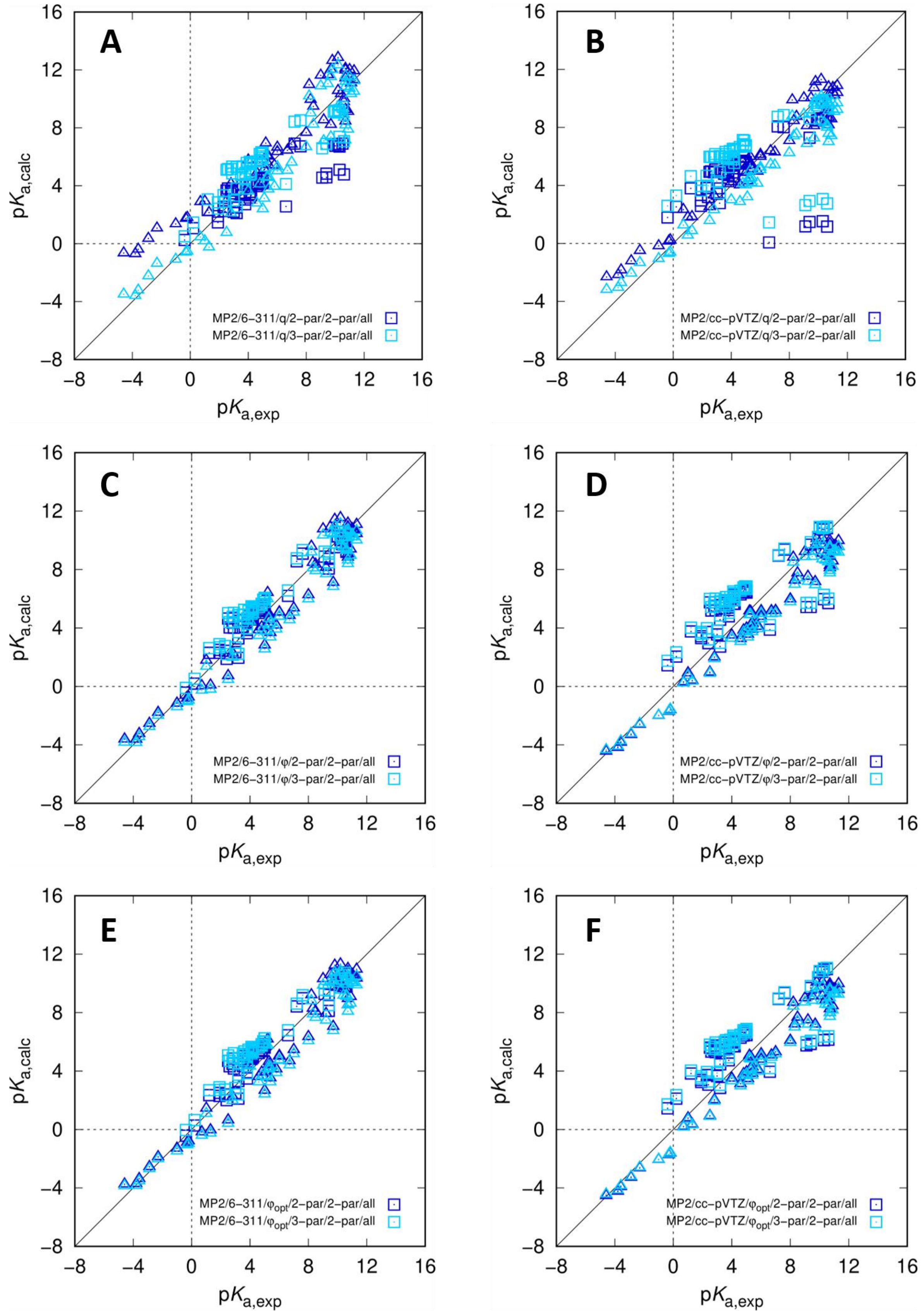


Fig. 9: Acidity constants pK_a calculated using EC-RISM at the MP2/6-311+G(d,p) (A,C,E) and MP2/cc-pVTZ (B,D,F) level of theory vs. the experimental results from the Klicić data set [143] using point charges (A,B), the exact electrostatic potential (C,D), or the periodicity-corrected exact electrostatic potential (E,F) with two parameters for the pK_a model. Data generated using the 2-par model for the PMV correction is shown in dark blue and data generated using the 3-par model as light blue. Acids are depicted as squares and bases as triangles. Raw

data are provided as part of OR_04, and optimized solution phase structures are provided as OR_05. The latter are identical for all models investigated in this chapter. Figure adapted from [2].

Table 9: Parameters *m* and *b* of optimized EC-RISM-based pK_a models and statistical metrics (root-mean-square error RMSE, mean absolute error MAE, and coefficient of determination R^2 from descriptive regression). Table adapted from [2].

EC-RISM	RMSE	MAE	*m*	*b*	R^2
MP2/6-311	nt/all	nt/all	nt/all	nt/all	nt/all
q/2-par/1-par	2.56/2.79	1.96/2.08	1/1	-205.973/-205.699	0.85/0.83
Acids	2.16/2.52	1.88/2.00			
Bases	2.87/3.03	2.03/2.16			
q/2-par/2-par	1.57/1.88	1.17/1.34	0.6369/0.6228	-129.239/-126.031	0.94/0.92
Acids	1.56/2.04	1.08/1.33			
Bases	1.57/1.70	1.26/1.34			
q/3-par/1-par	1.93/1.97	1.62/1.67	1/1	-204.923/-204.801	0.91/0.92
Acids	1.26/1.49	1.06/1.26			
Bases	2.37/2.34	2.12/2.08			
q/3-par/2-par	1.40/1.51	1.20/1.28	0.7332/0.7392	-148.488/-149.950	0.96/0.95
Acids	1.30/1.58	1.16/1.39			
Bases	1.48/1.43	1.23/1.18			
φ/2-par/1-par	1.64/1.61	1.37/1.32	1/1	-204.038/-204.057	0.94/0.94
Acids	1.12/1.08	0.91/0.87			
Bases	1.99/1.99	1.76/1.76			
φ/2-par/2-par	1.01/1.00	0.84/0.83	0.7438/0.7493	-150.393/-151.521	0.98/0.98
Acids	0.88/0.88	0.73/0.73			
Bases	1.12/1.11	0.94/0.93			
φ/3-par/1-par	1.79/1.77	1.54/1.50	1/1	-204.065/-204.112	0.93/0.93
Acids	1.39/1.33	1.21/1.16			
Bases	2.08/2.10	1.83/1.84			
φ/3-par/2-par	1.21/1.18	1.03/1.00	0.7363/0.7391	-148.832/-149.422	0.97/0.97
Acids	1.13/1.09	0.98/0.92			
Bases	1.27/1.26	1.08/1.07			
φ_{opt}/2-par/1-par	1.68/1.66	1.46/1.43	1/1	-204.197/-204.242	0.94/0.94
Acids	1.24/1.20	1.10/1.06			
Bases	1.99/2.00	1.77/1.78			
φ_{opt}/2-par/2-par	1.07/1.04	0.90/0.87	0.7421/0.7449	-150.155/-150.720	0.97/0.98
Acids	0.97/0.93	0.81/0.77			
Bases	1.15/1.14	0.97/0.97			
φ_{opt}/3-par/1-par	1.88/1.86	1.63/1.61	1/1	-204.221/-204.292	0.92/0.92
Acids	1.55/1.51	1.39/1.36			
Bases	2.12/2.15	1.84/1.87			
φ_{opt}/3-par/2-par	1.29/1.26	1.11/1.07	0.7290/0.7288	-147.428/-147.383	0.96/0.97
Acids	1.25/1.19	1.09/1.01			
Bases	1.33/1.33	1.13/1.13			
MP2/cc-pVTZ					
q/2-par/1-par	1.92/3.13	1.45/2.01	1/1	-207.786/-207.229	0.92/0.79
Acids	1.02/3.67	0.80/2.00			
Bases	2.45/2.50	2.03/2.01			
q/2-par/2-par	1.02/2.15	0.82/1.38	0.6973/0.5881	-143.263/-119.598	0.98/0.90
Acids	1.05/2.87	0.87/1.93			
Bases	1.00/1.06	0.78/0.84			
q/3-par/1-par	2.64/3.24	2.27/2.68	1/1	-207.012/-206.577	0.84/0.77
Acids	2.24/3.70	2.04/3.06			
Bases	2.95/2.72	2.47/2.31			
q/3-par/2-par	1.78/2.31	1.56/1.82	0.6395/0.5803	-130.455/-117.559	0.93/0.88
Acids	1.89/2.93	1.72/2.44			
Bases	1.68/1.50	1.43/1.23			
φ/2-par/1-par	2.40/2.51	2.00/2.13	1/1	-206.376/-206.176	0.87/0.86

Acids	2.04/2.45	1.85/2.20			
Bases	2.67/2.57	2.13/2.06			
φ/2-par/2-par	1.51/1.72	1.34/1.45	0.6579/0.6564	-133.937/-133.436	0.95/0.94
Acids	1.59/2.09	1.43/1.84			
Bases	1.44/1.27	1.25/1.08			
φ/3-par/1-par	2.65/2.71	2.23/2.32	1/1	-206.227/-206.050	0.84/0.84
Acids	2.40/2.67	2.27/2.45			
Bases	2.86/2.75	2.18/2.19			
φ/3-par/2-par	1.72/1.87	1.54/1.63	0.6333/0.6349	-128.634/-128.810	0.93/0.92
Acids	1.82/2.20	1.66/2.01			
Bases	1.64/1.48	1.44/1.27			
φ_{opt}/2-par/1-par	2.42/2.50	2.01/2.12	1/1	-206.333/-206.154	0.87/0.86
Acids	2.08/2.40	1.87/2.12			
Bases	2.68/1.70	2.13/1.46			
φ_{opt}/2-par/2-par	1.53/1.70	1.36/1.46	0.6560/0.6574	-133.507/-133.630	0.95/0.94
Acids	1.61/2.04	1.47/1.83			
Bases	1.45/1.30	1.27/1.11			
φ_{opt}/3-par/1-par	2.63/2.67	2.20/2.27	1/1	-206.212/-206.051	0.84/0.84
Acids	2.37/2.59	2.15/2.38			
Bases	2.84/2.74	2.25/2.18			
φ_{opt}/3-par/2-par	1.70/1.83	1.53/1.61	0.6357/0.6392	-129.135/-129.727	0.93/0.93
Acids	1.79/2.14	1.65/1.47			
Bases	1.61/1.47	1.43/1.27			

Looking at the results from calculations using the MP2/6-311/q/2-par/2-par/all PMV correction shows that the results are worse than those using the corresponding 3-par PMV correction that was originally used in the SAMPL5 challenge (see Fig. 9A). The RMSE is at 1.88 pK units for the entire data set, and acids are predicted significantly worse than bases with RMSEs of 2.04 and 1.70, respectively. This is in line with the results obtained for the MNSOL data set, where anions were predicted significantly worse than cations. The gap between acids and bases is decreased with RMSEs of only 1.58 and 1.43, yielding a total RMSE of 1.51 when the 3-par correction is used.

However, when applying the exact electrostatics during the EC-RISM calculations the results of the MP2/6-311/φ/2-par/2-par/all correction are actually slightly better than those using the corresponding 3-par correction, with total RMSEs of 1.00 and 1.18, respectively (see Fig. 9B). The improved electrostatics also invert the behavior of the acids and bases as in this case the acids show a lower error with an RMSE of only 0.88 compared to 1.11 for the bases. In all cases the results are greatly improved compared to the point charge model, with the best total RMSEs for the uncorrected and the periodicity corrected electrostatics at 1.00 (MP2/6-311/φ/2-par/2-par/all) and 1.04 (MP2/6-311/φ_{opt}/2-par/2-par/all), respectively, and when accounting for the statistical model errors these models are actually indistinguishable. While a convergence of the RMSEs for the 2-par and the 3-par models could be expected from the training data, a complete reversal was not seen on the MNSOL data set.

Comparing the results of the MP2/6-311+G(d,p) level of theory with those generated using the slightly larger cc-pVTZ basis set reveals unexpectedly bad results. Even when using exact, periodicity-corrected electrostatics, the best total RMSE of 1.70 for the MP2/cc-pVTZ/φ_{opt}/2-par/2-par/all model is not significantly better than that achieved by using the MP2/6-311/q/2-par/2-par/all model with atomic point charges and the smaller basis set. As expected, this is due to the bad performance of the anions that in this data set leads to a bad performance for the acids (RMSE 2.04) while the bases are predicted reasonably well (RMSE 1.30). Since only a single model is used for both acids and bases it is likely that the performance would be even better if the acids were not predicted as badly.

Especially for the point charge-based models MP2/6-311/q/*/2-par/all and MP2/cc-pVTZ/q/*/2-par/all the bad performance for thiols can be easily spotted in Fig. 9A and B, where significant outliers distort the predicted results. While the results are not as bad when using the full electrostatic potential, all models were also trained by removing the thiols from the training set, yielding the models MP2/6-311/*/*/*/nt and MP2/cc-pVTZ/*/*/*/nt. While the statistical metrics are incorporated in Table 9, the comparison with the experimental results is depicted in Fig. 10.

It is not surprising that the removal of the worst data points leads to improved statistical metrics, e.g. for the MP2/6-311/q/2-par/2-par/all and MP2/cc-pVTZ/q/2-par/2-par/all models the total RMSE shrinks from 1.88 and 2.15, respectively, to 1.57 and even 1.02, which is the best performance of any point charge model, for the corresponding “nt” models. This also represents the only case where the point charge-based model’s training set performance is better than the model’s using the exact electrostatic potential. In this case, the MP2/cc-pVTZ/φ_{opt}/2-par/2-par/nt yields an RMSE of 1.53, but even here the performance gap between the acids and bases shrinks significantly. Conversely, for the MP2/6-311-based models there is almost no difference in the performance when including or excluding the thiols from the training data set as long as the exact electrostatic potential is used.

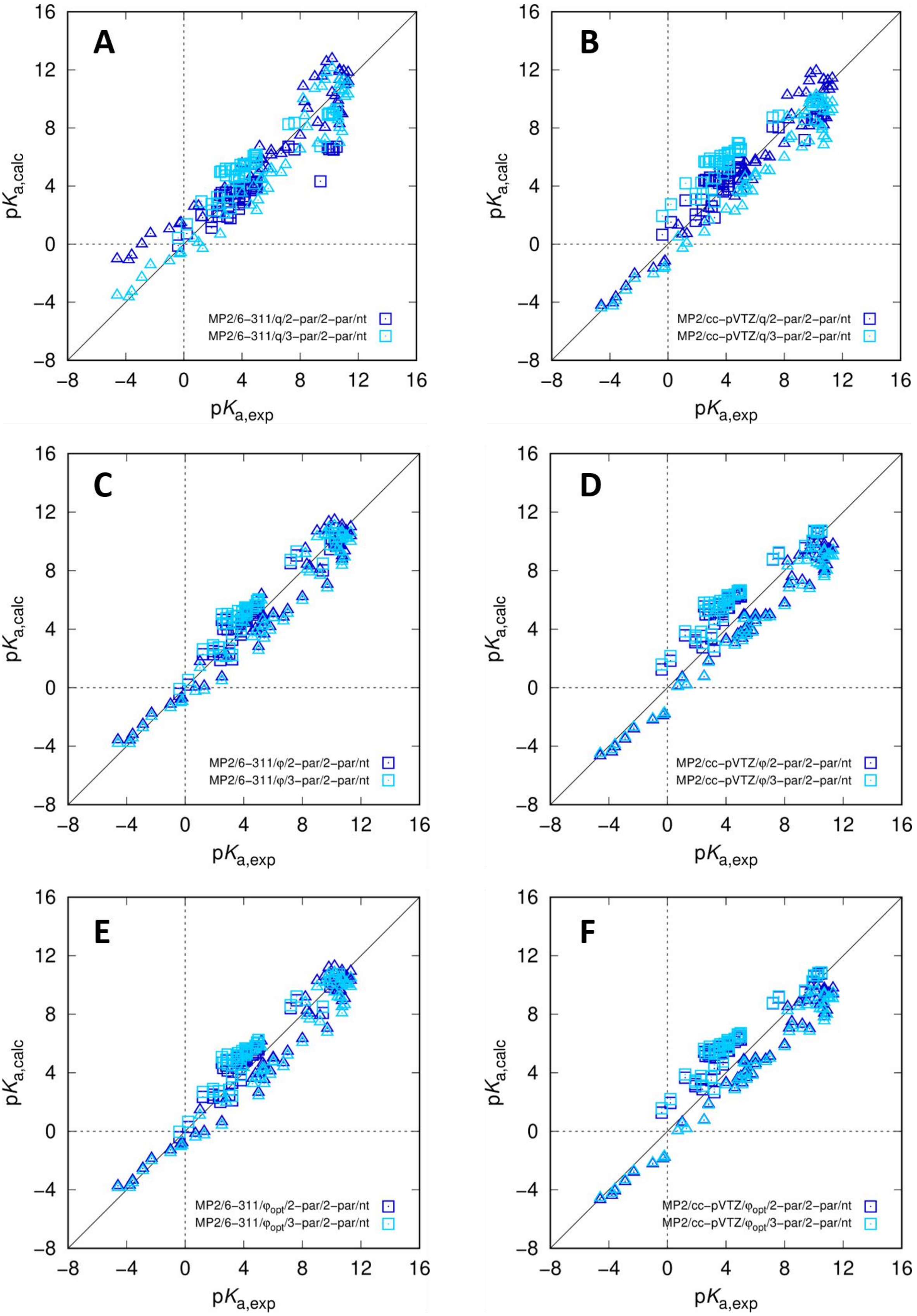


Fig. 10: Acidity constants pK_a calculated using EC-RISM at the MP2/6-311+G(d,p) (A,C,E) and MP2/cc-pVTZ (B,D,F) level of theory vs. the experimental results from the Klicić data set [143] using point charges (A,B), the exact electrostatic potential (C,D), or the periodicity-corrected exact electrostatic potential (E,F) with two parameters for the pK_a model. Thiols were excluded for the training of these models. Data generated using the 2-

par model for the PMV correction is shown in dark blue and data generated using the 3-par model as light blue. Acids are depicted as squares and bases as triangles. Raw data are provided as part of OR_04.

Using only a single parameter for the pK_a model unsurprisingly leads to worse results on the training data set, but here the differences in the errors between the models generally follow the differences for the pK_a model using two parameters. Again, the statistical metrics and model parameters are included in Table 9, while the comparison with the experimental values is depicted in Fig. 11.

It is easily apparent that the slope parameter is necessary for the accurate prediction of acidity constants from both a visual inspection of the predicted values in comparison with the experimental results and the statistical metrics: Looking at Fig. 11 shows that there are large differences between the experimental and the predicted acidity constants, especially for the smallest and largest values. Furthermore, the lowest values are underpredicted and the largest overpredicted, indicating that an additional slope parameter would improve the results. This is confirmed by the statistical metrices: Even the best-performing model MP2/6-311/φ/2-par/1-par/all only achieves an RMSE of 1.61 compared to 1.00 for the MP2/6-311/φ/2-par/2-par/all model. In general, the error is between 0.6 and 1.0 pK units bigger for all models when compared with their corresponding 2-par models and the smaller differences are exhibited by the better performing models using the exact electrostatic potential while the larger differences by the models using point charges.

These results indicate that for the MP2-based models the exact electrostatic potential should be used to predict accurate acidity constants. Furthermore, the 2-par PMV correction in combination with the 2-par pK_a model seem to generate the best results. The different models generated including or excluding thiols from the training set show only significantly differing results for the point charge-based models, but they still need to be tested on another data set to judge their relative performances.

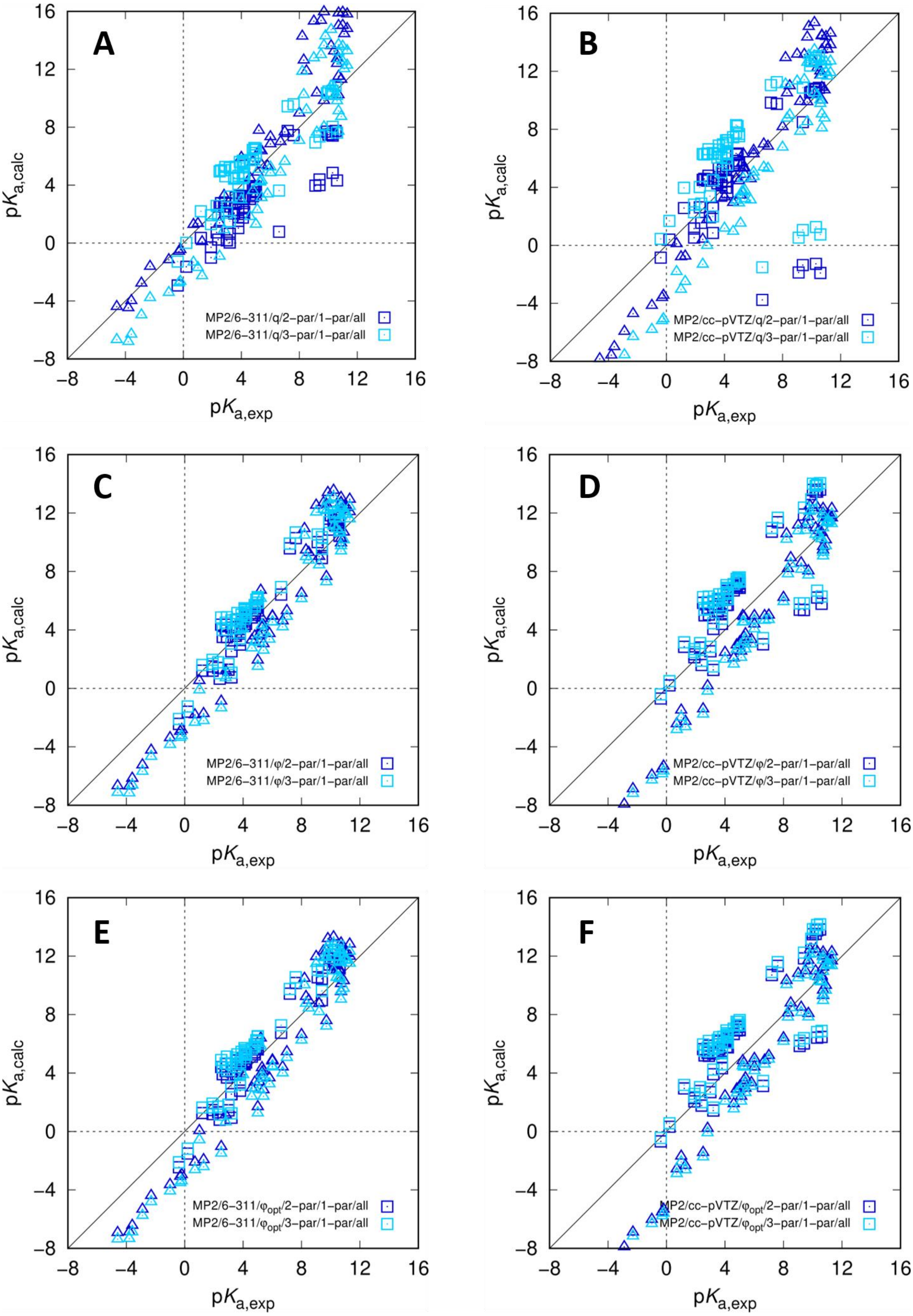


Fig. 11: Acidity constants pK_a calculated using EC-RISM at the MP2/6-311+G(d,p) (A,C,E) and MP2/cc-pVTZ (B,D,F) level of theory vs. the experimental results from the Klicić data set [143] using point charges (A,B), the exact electrostatic potential (C,D), or the periodicity-corrected exact electrostatic potential (E,F) with one parameters for the pK_a model. Data generated using the 2-par model for the PMV correction is shown in dark blue

and data generated using the 3-par model as light blue. Acids are depicted as squares and bases as triangles. Raw data are provided as part of OR_04.

5.3.2.2 DFT-based models

The less computationally expensive B3LYP/6-311+G(d,p) level of theory unsurprisingly exhibits worse results than what is achieved using MP2/6-311+G(d,p). With an RMSE of 2.71 for the 2-parameter point charge model, these results cannot be used to obtain accurate predictions. Here, the slope parameter is also smaller than 0.5 which indicates that the raw energy differences are the furthest from the true values when compared to the other two levels of theory. And while the 3-parameter PMV correction improves these results significantly, they are still worse than the corresponding MP2 models.

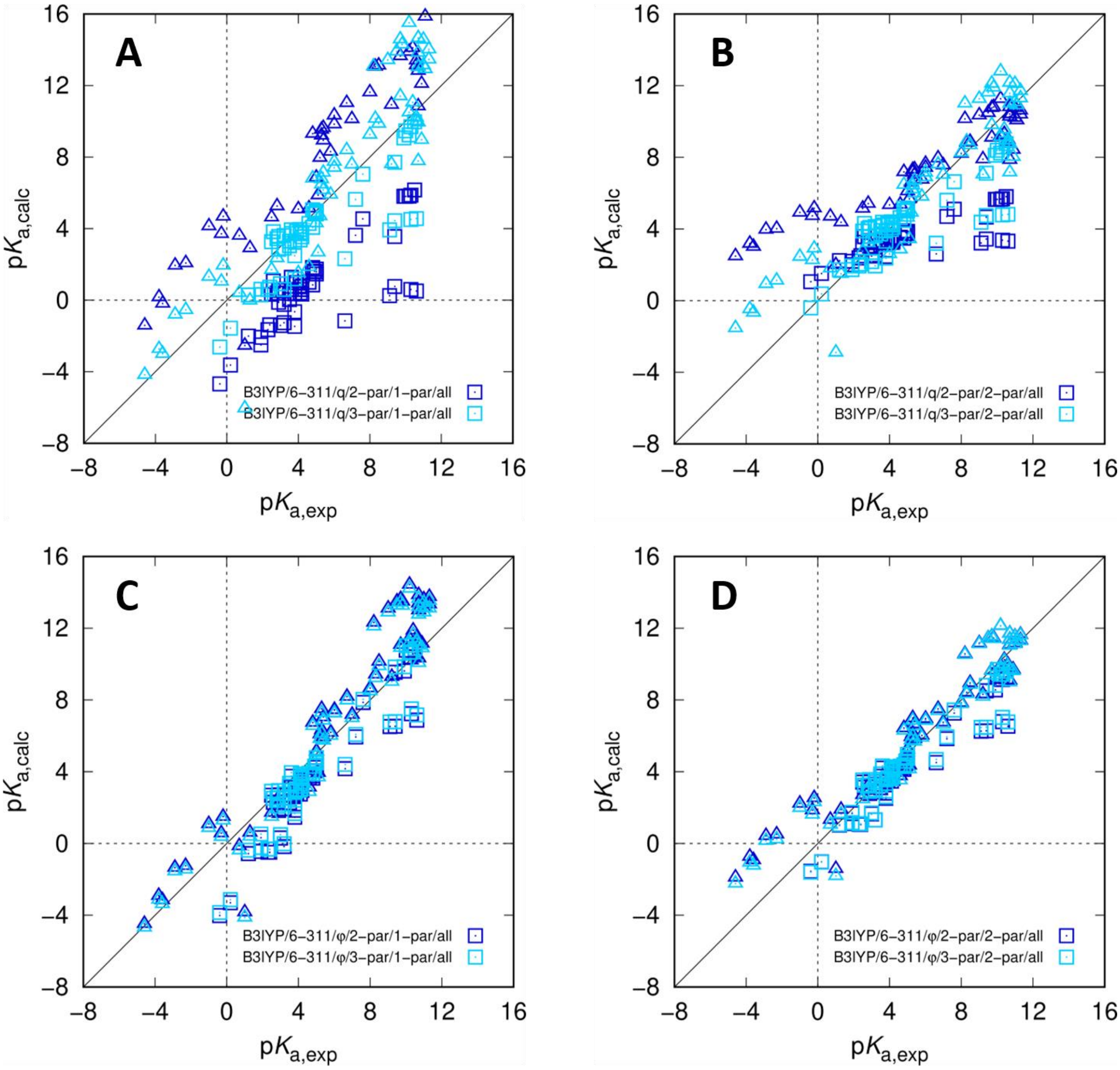


Fig. 12: Acidity constants pK_a calculated using EC-RISM at the B3LYP/6-311+G(d,p) level of theory vs. the experimental results from the Klicić data set using point charges (A,B) or the exact electrostatic potential (CD) with either a single parameter for the pK_a model (A,C) or two parameters (B,D). Data generated using two free

parameters for the PMV correction is shown in dark blue and data generated using three free parameters as light blue. Acids are depicted as squares and bases as triangles. Raw data are provided as part of OR_04.

Table 10: Parameters *m* and *b* of optimized EC-RISM-based pK_a models and statistical metrics (root-mean-square error RMSE, mean absolute error MAE, and coefficient of determination R^2 from descriptive regression).

Model	RMSE	MAE	*m*	*b*	R^2
EC-RISM					
B3LYP/6-311	nt/all	nt/all	nt/all	nt/all	nt/all
q/2-par/1-par	4.11/4.48	3.81/4.07	1/1	-206.252/-205.806	0.62/0.56
Acids	4.08/4.46	4.01/4.08			
Bases	4.13/4.51	3.65/4.06			
q/2-par/2-par	2.39/2.71	1.74/1.89	0.4789/0.4391	-95.987/-87.259	0.87/0.84
Acids	2.08/2.58	1.54/1.71			
Bases	2.64/2.83	1.92/2.07			
q/3-par/1-par	1.95/2.24	1.48/1.63	1/1	-204.915/-204.656	0.91/0.89
Acids	1.23/1.92	0.97/1.22			
Bases	2.41/2.50	1.91/2.01			
q/3-par/2-par	1.51/1.83	1.10/1.29	0.7433/0.7277	-150.946/-147.421	0.95/0.93
Acids	0.99/1.75	0.71/1.07			
Bases	1.84/1.90	1.45/1.51			
φ/2-par/1-par	1.79/1.86	1.40/1.47	1/1	-204.141/-203.995	0.93/0.92
Acids	1.59/1.68	1.25/1.32			
Bases	1.94/2.02	1.53/1.62			
φ/2-par/2-par	1.19/1.36	0.92/1.01	0.7361/0.7420	-148.864/-149.942	0.97/0.96
Acids	0.88/1.24	0.74/0.88			
Bases	1.41/1.48	1.07/1.14			
φ/3-par/1-par	1.66/1.73	1.27/1.33	1/1	-204.047/-203.915	0.94/0.93
Acids	1.43/1.53	1.09/1.16			
Bases	1.84/1.91	1.43/1.49			
φ/3-par/2-par	1.13/1.28	0.85/0.93	0.7523/0.7596	-152.185/-153.572	0.97/0.96
Acids	0.79/1.15	0.64/0.78			
Bases	1.35/1.40	1.03/1.08			
PCM/1-par/2-par	2.58	1.83	0.4022	-78.6463	0.85
Acids	2.09	1.49			
Bases	2.98	2.16			

Here, too, using the exact electrostatic potential improves the overall results, giving a model with a total RMSE of 1.28 for the best performing model using a 3-parameter correction, B3LYP/6-311/φ/3-par/2-par/all, and 1.36 for the best performing model using a 2-parameter correction, B3LYP/6-311/φ/2-par/2-par/all. Similar to the corresponding MP2/6-311 models, the acids are predicted more accurately than the bases, and if the small difference of only 0.36 in the RMSE would also be found in the independent test set it might be possible to use this much faster level of theory to calculate large molecules or datasets. The models trained by excluding all thiols perform slightly better than those including the thiols again, and the difference is exacerbated when using the point charge-based models. The largest difference amounts to 0.37 p*K* units for the B3LYP/6-311/q/2-par/1-par/nt model, while the smallest ones are only 0.07 for the B3LYP/6-311/φ/2-par/1-par/nt model, compared with the corresponding "all" models. As in the preceding chapter this variation of the model can only be discussed in more detail when applied to an independent test set. Due to time constraints during the challenge the periodicity-corrected exact electrostatic potential was not investigated further for the B3LYP/6-311+G(d,p) level of theory.

The acidity constants calculated with PCM at the B3LYP/6-311+G(d,p) level of theory are depicted in Fig. 13 and the corresponding parameters and statistical metrics for this method in Table 10. For the PCM model all molecules including thiols were used for training, because there was no significant deviation for these compounds. The results of this training are very questionable, especially for some of the acids and bases where large deviations between the experimental and predicted acidity constants can be found for many of the compounds with a low pK_a. Judging from the apparent individual slopes of the acids and bases, using separate models for the two compound classes better results might result in better predictions but this was not pursued during the challenge. While some compounds are predicted well it would be surprising if this model performed well on the SAMPL6 set of molecules because of the large number of outliers.

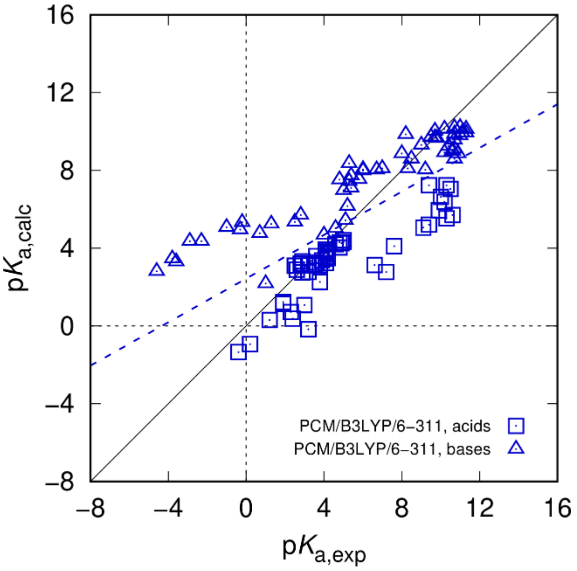


Fig. 13: Acidity constants pK_a calculated using PCM at the B3LYP/6-311+G(d,p) level of theory vs. the experimental results from the Klicić data set. Acids are depicted as squares and bases as triangles. Raw data are provided as part of OR_04.

5.3.3 pK_a model application

For the application of the models to the compounds of the SAMPL6 challenge a section dealing with the influence of ranking the conformations by their MM energies or their PCM energies is added to their designation. Here the use of the original MM conformations is denoted "c_{MM}", use of one PCM minimum structure is denoted "c_{QM}" and the use of the two lowest PCM structures is denoted as “c_{QM2}”. In the original publication these conformational ensembles were termed “c_{orig}”, “c_{opt}”, and “c_{opt2}”, respectively.

It should be noted in advance that while the SAMPL6 compounds can be considered a test set for the application of the models, it is by no means a perfect test set, because the chemical diversity differs significantly from that of both the MNSOL and the Klicić data set used to train the models. For example, all molecules in the SAMPL6 challenge contain multiple, substituted aromatic rings, while the compounds contained in the MNSOL range from simple aliphatic molecules, over simple alcohols, amines and other organic compounds to molecules that are comparable to the SAMPL6 challenge data set, e.g. caffeine, and even some inorganic ions. The statistical metrics for all models submitted during the SAMPL6 challenge and investigated in the post-submission phase are shown in Table 11.

Table 11: Statistical metrics for acidity constant predictions (root mean square error RMSE, mean absolute error MAE, mean signed error MSE, slope *m*’, intercept *b*’, and coefficient of determination R^2 from predictive regression) for the different model combinations submitted during the SAMPL6 challenge and generated in the post-submission phase. Models are sorted by their performance as measured by the RMSE. Raw and calculated data for both EC-RISM- and PCM-based models are provided as OR_06. Optimized structures are collected in OR_05.

EC-RISM	RMSE	MAE	MSE	*m*'	*b*’	R^2
MP2/6-311/φ_{opt}/2-par/2-par/all/c_{QM2}	1.13	0.97	-0.37	1.17	-1.38	0.91
MP2/6-311/φ_{opt}/2-par/2-par/all/c_{QM}	1.15	0.98	-0.39	1.16	-1.36	0.91
MP2/cc-pVTZ/φ_{opt}/2-par/2-par/all/c_{QM2}	1.15	1.01	-0.42	1.17	-1.45	0.91
MP2/cc-pVTZ/φ_{opt}/2-par/2-par/all/c_{QM}	1.23	1.09	-0.43	1.16	-1.41	0.90
MP2/6-311/φ_{opt}/2-par/2-par/all/c_{MM}	1.54	1.26	-0.37	1.22	-1.70	0.85
MP2/cc-pVTZ/φ/2-par/2-par/all/c_{MM}	1.60	1.24	0.32	1.14	-0.53	0.80
MP2/cc-pVTZ/q/2-par/2-par/all/c_{MM}	1.62	1.30	0.92	0.97	1.10	0.79
MP2/cc-pVTZ/φ/2-par/2-par/nt/c_{MM}	1.64	1.30	0.12	1.20	-1.08	0.81
MP2/6-311/φ/2-par/2-par/all/c_{MM}	1.70	1.24	-0.44	1.15	-1.36	0.79
MP2/cc-pVTZ/φ/3-par/2-par/all/c_{MM}	1.71	1.39	0.11	1.25	-1.43	0.82
MP2/cc-pVTZ/φ_{opt}/2-par/2-par/all/c_{MM}	1.71	1.40	-0.23	1.26	-1.82	0.83
MP2/6-311/φ/2-par/2-par/nt/c_{MM}	1.72	1.27	-0.51	1.15	-1.41	0.79
MP2-cc-pVTZ/q/2-par/2-par/nt/c_{MM}	1.80	1.39	0.74	1.15	-0.16	0.80
MP2/cc-pVTZ/φ/3-par/2-par/nt/c_{MM}	1.82	1.48	-0.10	1.29	-1.88	0.82
B3LYP/6-311/φ/2-par/2-par/nt/c_{MM}	1.99	1.56	0.59	1.01	0.50	0.67
MP2/6-311/φ/3-par/2-par/all/c_{MM}	2.01	1.59	-0.52	1.36	-2.68	0.82
MP2/6-311/φ/3-par/2-par/nt/c_{MM}	2.01	1.58	-0.56	1.35	-2.66	0.82
MP2/cc-pVTZ/q/3-par/2-par/all/c_{MM}	2.10	1.69	0.36	1.40	-2.06	0.82
B3LYP/6-311/φ/3-par/2-par/nt/c_{MM}	2.21	1.65	0.73	1.28	-0.98	0.76
MP2/6-311/q/3-par/2-par/nt/c_{MM}	2.22	1.78	-0.78	1.41	-3.24	0.82
MP2/6-311/φ/2-par/1-par/all/c_{MM}	2.40	1.94	-0.31	1.55	-3.64	0.84
MP2/cc-pVTZ/q/3-par/2-par/nt/c_{MM}	2.44	2.06	0.11	1.54	-3.18	0.82
B3LYP/6-311)/q/3-par/2-par/nt/c_{MM}	2.54	1.83	0.65	1.43	-1.96	0.76
PCM/B3LYP/6-311/c_{MM}	2.84	2.63	0.57	0.08	6.12	0.03
MP2/6-311/φ/3-par/1-par/all/c_{MM}	2.99	2.53	-0.42	1.78	-5.15	0.84

The results of the SAMPL6 challenge predictions partially exhibits similar trends as the training set, but there are some noticeable differences. The B3LYP energies give significantly worse predictions for the pK_a than the two levels of theory that use MP2 energies when using point charges. A comparison of the predicted acidity constants for the three different B3LYP models using either point charges or the full electrostatics without the periodicity correction, which was only developed during the post-submission phase, are shown in Fig. 14A. This gap is partially closed when using the exact electrostatic potential, but the MP2 models still perfom better, because even the best B3LYP model, B3LYP/6-311+G(d,p)/φ/2-par/2-par/nt/c_{MM} only achieves an RMSE of 1.99 while the corresponding MP2 model has an RMSE of 1.70. Despite using an additional parameter for the PMV correction the other two B3LYP-based models show an even worse performance with RMSEs of 2.21 and 2.54, indicating that the additional parameter actually worsens the model performance.

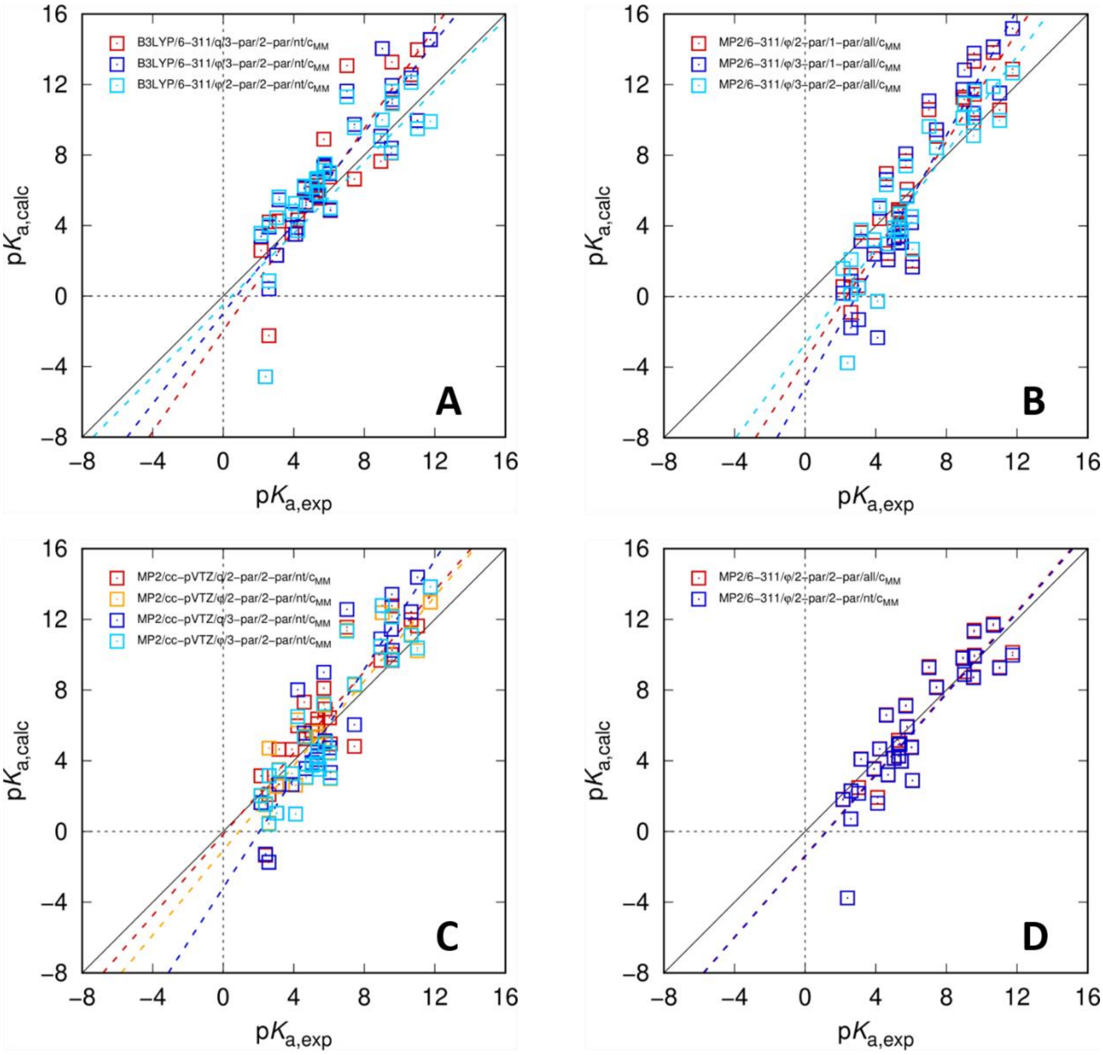


Fig. 14: Acidity constants pK_a calculated using EC-RISM at the B3LYP/6-311+G(d,p) (A), MP2/6-311+G(d,p) (B,D), and MP2/cc-pVTZ (C) level of theory using only the original c_{MM} structures vs. the experimental results from the SAMPL6 challenge data set.

For the MP2/6-311 models the same trend can be observed. Compared to the aforementioned MP2 model with an RMSE of 1.70, the corresponding 3-par model MP2/6-311/φ/3-par/2-par/all/c_{MM} yields an RMSE of 2.01. Most noticeably the slope of the descriptive regression is 1.15 for the 2-par and 1.36 for the 3-par model, which shows that the deviation from the ideal slope is more than twice as big (see Fig. 14B and D). A similar trend is also seen in Fig. 14C for models using the cc-pVTZ basis set. The 3-par model using the exact electrostatic potential barely achieves the same performance as the 2-par model using point charges, with RMSEs of 1.82 and 1.80, respectively, and both the slope and the offset of the descriptive regression are significantly worse. The best model of this series of related models is the MP2/cc-pVTZ/φ/2-par/2-par/nt/c_{MM} model with an RMSE of 1.64. Also shown in Fig. 14D is confirmation that the exclusion of thiolic compounds makes no meaningful difference for models using the 6-311+G(d,p) basis set and the exact electrostatic potential. This could be expected, because the model parameters are almost identical for these models and the RMSEs of 1.70 and 1.72 for the MP2/6-311/φ/2-par/2-par/all/c_{MM} and MP2/6-311/φ/2-par/2-par/nt/c_{MM} models, respectively or identical RMSEs of 2.01 for both the MP2/6-311/φ/3-par/2-par/all/c_{MM} and MP2/6-311/φ/3-par/2-par/nt/c_{MM} models confirm this.

Finally, the models using a 1-par pK_a model generally perform worse than their corresponding 2-par models. An example of this is depicted in Fig. 14B where the MP2/6-311/φ/3-par/2-par/all/c_{MM} model with an RMSE of 2.01 is still significantly better than the 1-par model with an RMSE of 2.99. The descriptive regression slopes of 1.36 for the 2-par model compared to 1.78 for the 1-par model all but confirms the necessity of using two parameters to achieve reasonably accurate performance. The same observation can be made for the MP2/6-311/φ/2-par/1-par/all/c_{MM} model (RMSE: 2.40, *m*': 1.55), where the comparable MP2/6-311/φ/2-par/2-par/all/c_{MM} model outperforms it in a similar manner (RMSE: 1.70, *m*': 1.15).

Looking at the best-performing models with regards to the RMSE in Table 11 it is noticeable that only models using the periodicity corrected electrostatic potential are at the top of the ranking and there is a gap of 0.47 in the RMSE between the best-performing model and the first model using the normal electrostatic potential. The best model, MP2/6-311/$φ_{opt}$/2-par/2-par/all/c_{QM2}, achieves an RMSE of only 1.13, and despite the worse performance of the cc-pVTZ basis set of acids during the training set, the corresponding MP2/cc-pVTZ/$φ_{opt}$/2-par/2-par/all/c_{QM2} model is only 0.02 p*K* units worse.

An important question to consider is, how much of this improved performance is due to the periodicity correction (q, φ, and $φ_{opt}$), and how much of it is due to the different conformations used in the EC-RISM calculations (c_{MM}, c_{QM}, and c_{QM2}). These comparisons are depicted in Fig. 15.

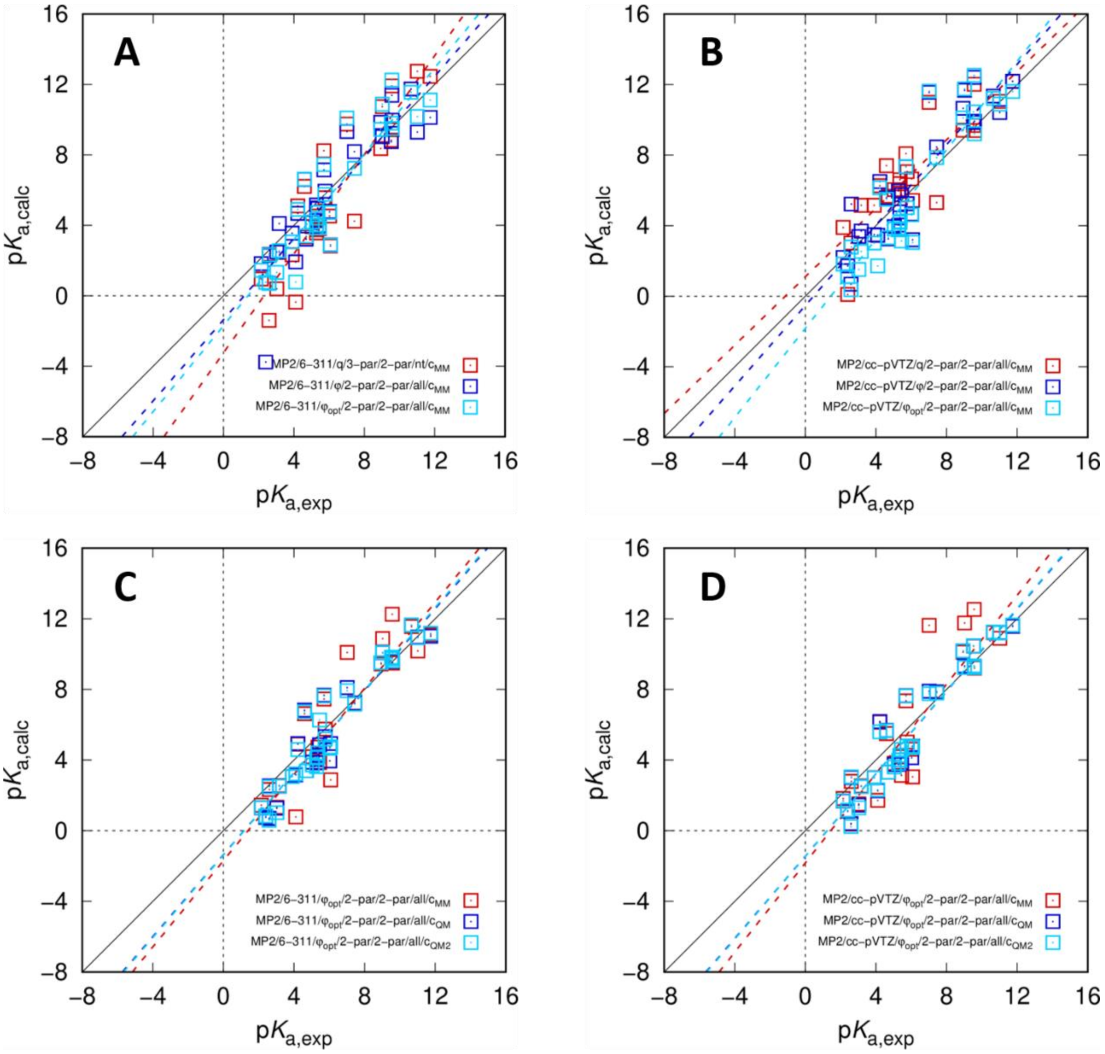


Fig. 15: Acidity constants pK_a calculated using EC-RISM at the MP2 level of theory using the 6-311+G(d,p) (A,C) and the cc-pVTZ basis sets (B,D) vs. the experimental results from the SAMPL6 challenge data set. In panel A the model switches more significantly because the exact corresponding model was not investigated during the SAMPL6 challenge.

The results obtained from the original point charge-based models and those using exact electrostatics or even the periodicity corrected implementation does not appear to differ significantly. A series of MP2/cc-pVTZ/*/2-par/2-par/all/c_{MM} models is best to illustrate this (Fig. 15B): The point charge-based model of this series yields an RMSE of 1.62, while the φ and $φ_{opt}$ models give an RMSE of 1.60 and 1.71, respectively. In this case the use of the periodicity correction increases the RMSE slightly. For the models based on the series denoted as MP2/6-311/*/2-par/2-par/all/c_{MM} this comparison can not be made fully, because this point charge-based model was not investigated, but the results for the exact electrostatic potential indicate an improvement from an RMSE of 1.70 to 1.64 when using the periodicity correction (Fig. 15A). In any case, the periodicity correction is nevertheless very beneficial because for no significant computational overhead the number of EC-RISM calculations diverging due to constantly increasing polarization of the molecule was reduced to zero when using this implementation and even for the models using the cc-pVTZ basis set the RMSEs are very close. It must also be considered that there might be some error compensation resulting from the use of c_{MM} conformations in combination with the q- or φ-models.

This is due to the large contribution to the model error caused by the choice of the conformations, which is depicted in Fig. 15C and D. Early in the challenge only the conformation with the lowest MM-energy was further optimized for each microstate and their energy calculated with EC-RISM. While these conformations are optimized quantum-chemically, the local minimum found by that optimization is not necessarily close to the global minimum, because many intramolecular interactions are not captured by the MM-energy evaluation. Upon re-optimization of the entire set of conformations and re-ranking them by their PCM energy, the EC-RISM results obtained from the minimum structure significantly improve the predicted pK_a values. The results obtained by using the MP2/6-311/φ_{opt}/2-par/2-par/all/c_{MM} model show an RMSE of 1.64 with reasonable slope parameters of 1.22 from the descriptive regression. Compared to that, just by using the c_{QM} conformations the RMSE is improved significantly to 1.15 and the slope parameter slightly to 1.17. the Using the second-lowest conformation here barely improves the results any further to an RMSE of 1.13, indicating that just the single minimum QM structure is one of the most important factors to predict accurate values for the pK_a. The results of the corresponding models using the cc-pVTZ basis set, MP2/6-311/φ_{opt}/2-par/2-par/all/*, support this conclusion, too. Here, the RMSE improves from 1.71 for the c_{MM}-model, to 1.23 for the c_{QM}-model to 1.15 for the c_{QM2}-model, again showing the importance of the minimum QM structures.

All these results point to the MP2/6-311/φ_{opt}/2-par/2-par/all/c_{QM2} model as the best practice for predicting pK_a values. While the MP2/cc-pVTZ/φ_{opt}/2-par/2-par/all/c_{QM2} model has a similar performance, this comes at a significantly increased computational cost.

One issue that might have occurred for all models considered up to this point during the prediction of acidity constants for the SAMPL6 data set is the calculation of Gibbs energies for species with a charge greater than 1 or smaller than -1. Neither the MNSOL nor the Klicić data set contains multiply charged molecules, so if the effect of the unphysical process corrected for by the c_q parameter of the PMV correction does not scale linearly with the charge, larger deviations from the experimental values can be expected for these acids and bases. Fortunately, the acidity constants of the compound SM18 are predicted exceptionally well, showing that the PMV correction and the pK_a models originally trained solely on singularly charged compounds are applicable even outside the chemical space they were trained in. Predicted and experimental acidity constants for the individual compounds of the SAMPL6 challenge are shown in Table 12 for a few selected models.

Table 12: Individual experimental and assigned computational data for macrostate pK_a values for various models based on MP2/6-311/*/2-par/2-par/all/*. Red text color indicates data points where solutions with exact electrostatic could not be obtained (original submission only) and were replaced by point charge data. Table adapted from [2].

	p$K_{a,exp}$	φ_{orig}/c_{orig}	φ_{opt}/c_{orig}	φ_{opt}/c_{opt}	φ_{opt}/c_{opt2}
	Exp.	MP2/6-311+G(d,p)			
SM01	9.53	8.75	9.81	9.81	9.82
SM02	5.03	4.16	4.18	3.88	3.73
SM03	7.02	9.32	10.10	8.10	7.94
SM04	6.02	4.78	4.74	3.95	4.67
SM05	4.59	6.61	6.62	6.83	6.74
SM06	3.03	2.48	1.32	1.27	1.02
	11.74	10.12	11.12	11.02	11.19
SM07	6.08	2.90	2.88	4.93	4.75
SM08	4.22	4.68	4.90	4.94	4.60
SM09	5.37	4.95	4.74	4.27	4.17
SM10	9.02	9.04	10.87	10.07	10.10
SM11	3.89	3.55	3.06	3.06	3.06
SM12	5.28	5.15	3.84	3.84	3.62
SM13	5.77	5.94	5.76	5.32	5.19
SM14	2.58	0.72	0.70	0.69	0.59
	5.30	4.28	4.02	4.03	4.13
SM15	4.70	3.21	3.38	3.38	3.38
	8.94	9.85	9.44	9.52	9.52
SM16	5.37	4.99	4.40	4.39	4.41
	10.65	11.74	11.56	11.64	11.64
SM17	3.16	4.11	2.55	2.52	2.52
SM18	2.15	1.82	1.44	1.29	1.31
	9.58	9.97	9.49	9.59	9.62
	11.02	9.30	10.18	10.97	10.92
SM19	9.56	11.38	12.25	9.80	9.74
SM20	5.70	7.15	7.45	7.68	7.63
SM21	4.10	1.92	0.78	3.16	3.22
SM22	2.40	-3.77	0.77	0.77	0.83
	7.43	8.19	7.23	7.22	7.16
SM23	5.45	3.98	3.87	4.87	6.26
SM24	2.60	2.30	2.33	2.53	2.48

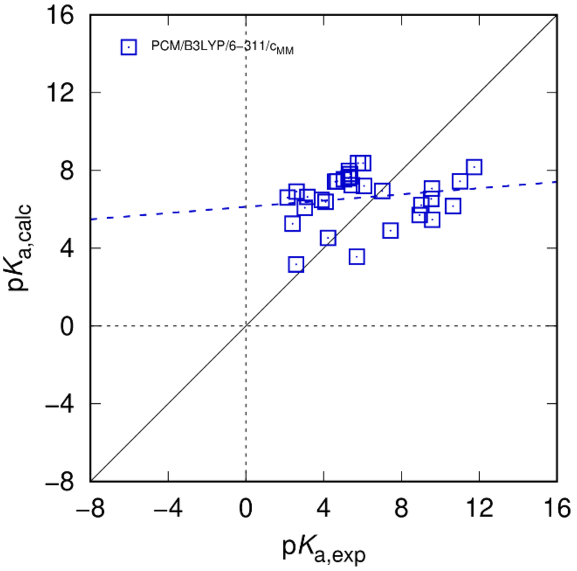


Fig. 16: Acidity constants pK_a calculated using PCM energies at the B3LYP/6-311+G(d,p) level of theory and corrected with the PCM/B3LYP/6-311/c$_{MM}$ model vs. the experimental results from the SAMPL6 challenge data set.

The acidity constants of the SAMPL6 dataset calculated with PCM at the B3LYP/6-311+G(d,p) level of theory and applying the PCM/B3LYP/6-311/c$_{MM}$ model are depicted in Fig. 16. This model's performance is singular in that it has not just a high RMSE of 2.84, which is one of the worst of the models investigated here, but the regression parameters are complete outliers compared to all other models as well. With a slope of 0.03 and an offset of 6.12 there is very little correlation between the predicted and the experimental value. The reason for the RMSE not being even higher is the small dynamic range of the experimental values that range from 2.40 to 11.74 with most compounds tending towards the middle of this range. The performance of a possible null hypothesis such as "every compound has a pK_a of 7" is comparable, yielding an RMSE of 2.86. This implies that the predictive power of the PCM model as implemented here is close to zero and it cannot even be used as a method to get a quick estimate for a compound's pK_a.

5.3.4 Prediction of pH-dependent tautomer ratios

The additional task of predicting the pH-dependent tautomer ratios could be easily completed using just the individual Gibbs energies of the different tautomers in their individual protonation states, but initially no experimental results were available for comparison. This was achieved by applying eq. (59) originally derived by L. Eberlein to the calculated EC-RISM data.

After the SAMPL6 challenge the organizers were able to experimentally determine the microstates of the compounds SM07 and SM14 in the two ionization states, 0 and +1 through analysis of nuclear magnetic resonance (NMR) parameters [153]. For this compound the microstate SM07_micro004 was determined to be exclusively present for the neutral state and the microstate SM07_micro006 for the protonated state. While the absolute pK_a for this compound was not among the best predicted ones, with a deviation of 1.33 for the best-performing MP2/6-311+G(d,p) model, the microstates were nevertheless predicted correctly for both protonation states. The microstate SM07_micro004 was predicted to account for 0.996 of the total population, which would make the other tautomers experimentally undetectable, and the microstate SM07_micro006 was predicted with an even larger population. For the compound SM14 that had two titratable sites within the experimental range the tautomers SM14_micro001, SM14_micro002 and SM14_micro003 were determined to be exclusively present in the neutral, protonated, and doubly protonated state, respectively. Again, the predicted pK_a values are not ideal, with deviations of 1.89 and 1.27, but the microstates predicted using EC-RISM for the neutral and the doubly protonated ionization states are the correct one and all other microstates in these protonation states are correctly predicted to not contribute. For the ionization state with only a single protonation the NMR analysis suggests only the existence of the microstate SM14_micro002, the predicted microstates yield a relative population of 0.621 for SM14_micro004 and only 0.379 for SM14_micro002.

An interesting case is the compound SM13. While no experimental NMR data was produced for this compound it is structurally similar to SM07, with a phenyl substituting the benzyl group and two additional methoxy substituents. The structures of both compounds are depicted in Fig. 17.

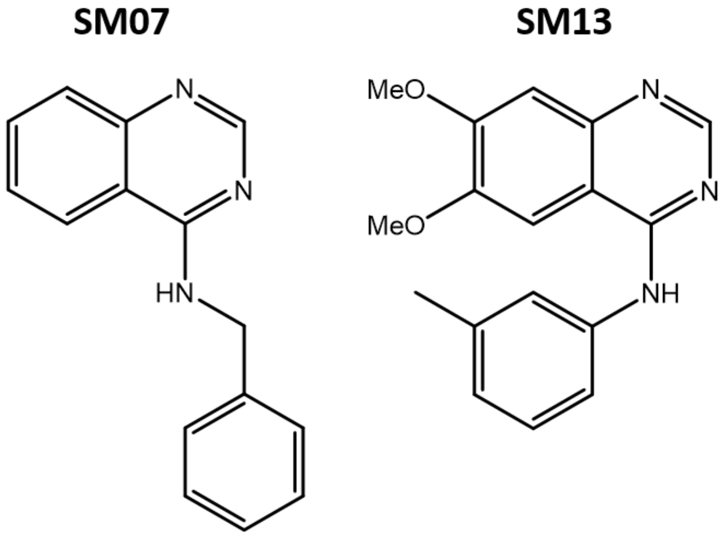


Fig. 17: Chemical structures of the tautomerically stable compound SM07 and the similar compound SM13 that is predicted to have two different tautomers in its protonated form by EC-RISM.

Despite this structural similarity, EC-RISM predicts the protonated states to be made up of two relevant microstates with populations of 0.626 and 0.374, respectively. The different microstates and their total populations as a function of the pH are shown in Fig. 18. For this compound the predicted pK_a is close to the experimental value as well with a deviation of only 0.58, implying that the EC-RISM energies are reasonable. The tautomeric analogue of this second state could also exist for SM07 but is there correctly predicted to be so energetically unfavorable that it cannot be detected.

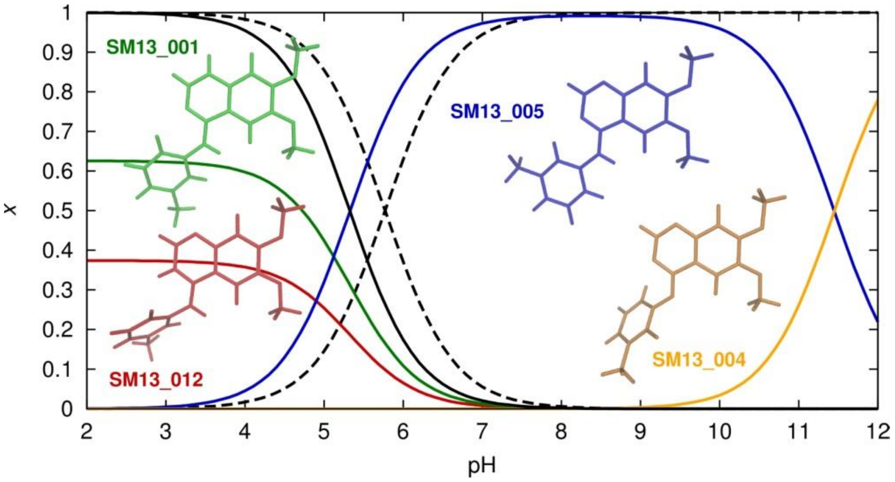


Fig. 18: Exemplary pH-dependent population for all relevant microstates of compound SM13 (shown in their respective optimal conformations) from EC-RISM calculations (MP2/6-311/φ_{opt}/2-par/2-par/all/c_{QM2}), and experimental and theoretical macrostate populations. Dashed lines: from experimental pK_a value; solid lines: calculation results. Non-black lines represent microstates; black line: macrostate populations; since only one microstate is predicted for the neutral and anionic macrostate, the blue and orange line also represent macrostate populations. Material from: 'N Tielker, L Eberlein, S Güssregen, SM Kast, The SAMPL6 challenge on predicting aqueous pK_a values from EC-RISM theory, J Comput-Aided Mol Des, published 2018, Springer' [2].

Unfortunately, the absence of experimental results for this compound makes confirmation or disproval of this prediction impossible for the time being, but further investigations of experimentally determined tautomer ratios might be interesting, especially in light of the good performance of EC-RISM in the SAMPL2 challenge where relative tautomer stabilities in water had to be determined [43].

6 SAMPL6.2: PREDICTION OF PARTITION COEFFICIENTS BETWEEN WATER AND OCTANOL

6.1 Introduction

The second part of the SAMPL6 challenge dealt with predicting the neutral state partitioning behavior between water and octanol. This organic phase is in some regards simpler than the cyclohexane of the SAMPL5 challenge, for example both Gibbs energies of solvation and especially Gibbs energies of transfer between this solvent and water are far more abundant in the literature than they are for cyclohexane. The reason for this is the fact that the octanol-water partition coefficient has long been used to predict the lipophilicity of compounds and estimate their permeability in the human body [7]. On the other hand, the specific chemical makeup of octanol makes it more difficult to handle for modeling purposes. The hydroxyl group gives this molecule a certain polarization, meaning that EC-RISM is once again necessary to capture the induced polarization of the compounds upon solvation, but more significantly the amount of water contained in the octanol phase might be non-negligible. That is why the best performing water model of the SAMPL6.1 challenge was reused here, while the only thing that was varied was the octanol model. The PMV correction was trained using a single and two parameters, once for dry octanol and once for the octanol-water mixture with the experimental density and molar fraction of water, leading to a total of four models under consideration. Here the corrected excess chemical potentials for the different models are

$$\mu_{\mathrm{corr}}^{\mathrm{ex}} = c_{\mu}\mu^{\mathrm{ex}} + c_{v}V^{m} \tag{79}$$

for the 2-par model and

$$\mu_{\mathrm{corr}}^{\mathrm{ex}} = \mu^{\mathrm{ex}} + c_{v}V^{m} \tag{80}$$

for the 1-par model.

In addition, the training function for compounds with more than a single conformation was changed slightly. For this PMV correction the properly weighted target function

$$\{c_\mu, c_V\} = \arg\min\left[\sum_{\text{molecules}}\left(-\beta^{-1}\ln\sum_{tc}\exp[-\beta(E_{tc}^{\text{sol}}+\mu_{\text{corr}}^{\text{ex}})]-\beta^{-1}\ln\sum_{tc'}\exp[-\beta E_{tc'}^{\text{vac}}]-\Delta_{\text{solv}}G_{\text{exp}}^{0}\right)^2\right] \tag{81}$$

is used, where t and c represent the tautomers and conformations, respectively instead of using only the minimum conformation for training the parameters.

6.2 Computational details

The conformations generated for the SAMPL6.1 challenge were used unchanged for the water phase and reoptimized for the octanol phase at the IEFPCM/B3LYP/6-311+G(d,p) level of theory using Gaussian 16 rev. B.01 [154] with tight convergence criteria and octanol as a PCM solvent. Due to the fact that charged species did not have to be evaluated in this part of the challenge, every generated conformation could be treated with EC-RISM for both solvents instead of only the two lowest energy conformations. In this case the PMV was calculated via the total correlation route [76,77], using the 1D RISM estimate of the isothermal compressibility with a value of $0.717062\cdot10^{-9}$ Pa^{-1} for water and the experimental compressibility of $0.761\cdot10^{-9}$ Pa^{-1} for octanol [155]. Due to convergence issues, for octanol the PSE-1 closure was used throughout, while all other 3D RISM and EC-RISM settings were taken unchanged from the earlier part of the SAMPL6 challenge.

6.3 Results

6.3.1 Solvation free energies

While the water model from the SAMPL6.1 challenge could be used unchanged, the octanol PMV correction had to be created from scratch. The MNSOL contains 224 molecules with experimental Gibbs energies of solvation in octanol. Training the correction was conducted using either only one parameter for the partial molar volume or an additional parameter directly scaling the excess chemical potential, because the effect of this had not yet been investigated for octanol in the same way it had been for water and cyclohexane. Results for the MNSOL compounds without the PMV correction and for the two different PMV corrections for both the dry and wet octanol solvent model are shown in Fig. 19 while the parameters and the statistical metrics are shown in Table 13.

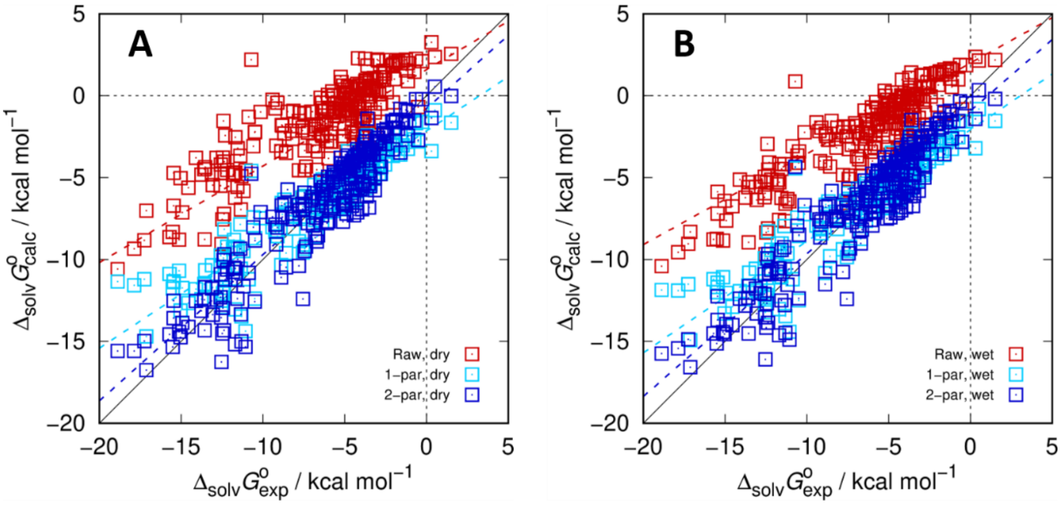


Fig. 19: Gibbs energies of solvation in octanol calculated using EC-RISM vs. the experimental results from the MNSOL database for both the dry (A) and the wet (B) models. Uncorrected data is shown using red squares, the data generated using the 1-par correction is shown in light and that generated using the 2-par correction in dark blue. Raw data are provided as part of OR_07. Optimized solution and gas phase structures are provided as OR_08. Figure adapted from [5].

Table 13: Regression parameters of optimized EC-RISM-based Gibbs energy of solvation models (c_q, c_V / kcal mol^{-1} Å^{-3}, c_q / kcal mol^{-1} e^{-1}) along with statistical metrics (root-mean-square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*', intercept *b*' / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression). c_V corresponds to PMVs computed using the total correlation function with an experimental isothermal compressibility of 0.761·10^{-9} Pa^{-1} for octanol and the RISM estimate of 0.717062·10^{-9} Pa^{-1} for water. Material from: 'N Tielker, D Tomazic, L Eberlein, S Güssregen, SM Kast, The SAMPL6 challenge on predicting octanol–water partition coefficients from EC-RISM theory, J Comput-Aided Mol Des, published 2020, Springer' [5].

Solvent	RMSE	MAE	MSE	*m*'	*b*'	R^2	c_μ	c_V	c_q
Water									
All	2.04	1.43	-0.26	1.00	-0.35	1.00	-	-0.10251	-15.728
Neutrals	1.56	1.13	-0.36	0.97	-0.47	0.89	-	-	-
Anions	3.07	2.46	0.01	1.10	7.18	0.94	-	-	-
Cations	2.98	2.10	0.02	0.96	-2.62	0.85	-	-	-
Octanol (dry)									
1-par	1.78	1.33	0.03	0.66	-2.15	0.85	-	-0.00799	-
2-par	1.48	1.14	-0.08	0.89	-0.78	0.87	1.33446	-0.00609	-
Octanol (wet)									
1-par	1.73	1.31	-0.01	0.68	-2.08	0.85	-	-0.01552	-
2-par	1.51	1.16	-0.10	0.87	-0.93	0.86	1.28924	-0.01315	-

Regression results for the PMV correction of octanol exhibit similar trends as the other solvents investigated in this work. One noteworthy aspect is that the uncorrected Gibbs energies of solvation are much closer to the experimental values than for the other two solvents investigated in this work. This is also reflected in the c_V parameter that is one order of magnitude smaller than that for water. Part of the reason for the smaller parameter is the larger partial molar volumes calculated by 3D RISM. However, the magnitude of the difference cannot be explained solely by this.

There is seemingly very little difference between the dry and the wet octanol model despite the significant molar fraction of water in the latter model. While there is some deviation in the correction parameters, all the statistical metrics are within the margin of error between the two models. As in the case of cyclohexane the model with an additional parameter, in this case the 2-par model, performs slightly better than the 1-par model. But as the experience in the SAMPL5 challenge shows this should not be taken as proof that the 2-par model is superior because overfitting can occur. This is one aspect that makes the SAMPL challenges so valuable, as they make it possible to investigate potential model errors.

6.3.2 MNSOL partition coefficients

The MNSOL database contains not only Gibbs energies of solvation, but also Gibbs energies of transfer between two solvents and the solvent pair octanol-water is among them. These Gibbs energies of transfer can be easily converted into partition coefficients log *P* and this serves as a first, but not truly independent test set, because the molecules in this data set are merely a subset of the Gibbs energies of solvation in octanol, which are in turn a subset of the Gibbs energies of solvation in water that were used to train the PMV corrections. Still, this can provide a first glimpse at the relative performance of the four different models, so the predicted and experimental partition coefficients for the MNSOL compounds are depicted in Fig. 20 while the statistical parameters for the different octanol models are shown in Table 14.

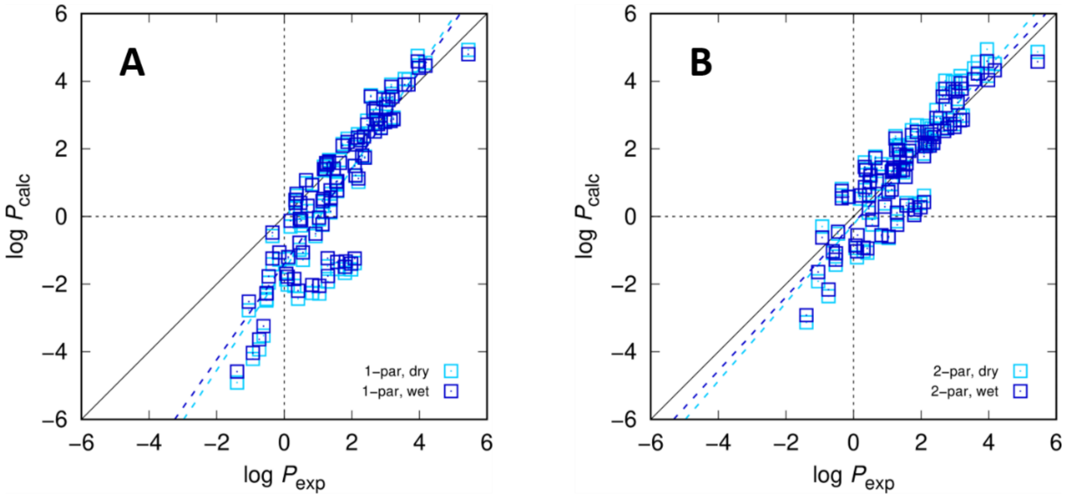


Fig. 20: Partition coefficients log *P* between water and octanol calculated using EC-RISM vs. the experimental values calculated from the Gibbs energies of transfer according to Eq. (60) in the MNSOL database for both the 1-par (A) and the 2-par correction (B). Data generated using the dry octanol model is shown in light blue while that generated using the wet octanol model is shown in dark blue. Raw data are provided as part of OR_07 and structures as OR_08.

Table 14: Statistical metrics for log *P* predictions on the MNSOL transfer free energy data set (root-mean-square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*', intercept *b*' / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression).

Solvent	RMSE	MAE	MSE	*m*'	*b*'	R^2
Octanol (dry)						
1-par	1.51	1.08	-0.83	1.49	-1.57	0.77
2-par	0.83	0.68	0.01	1.16	-0.23	0.79

Octanol (wet)						
1-par	1.40	0.99	-0.77	1.42	-1.40	0.77
2-par	0.77	0.61	-0.06	1.09	-0.20	0.79

The results show two important trends. Firstly, the 2-par corrections are significantly better at predicting the correct partition coefficients. The worse results of the 1-par corrections are not caused by a larger spread of the predicted values but by a significant underestimation of the partition coefficients that increases even more with lower log *P* values. This indicates that the parameter directly scaling the excess chemical potential may be necessary to correctly model octanol as a solvent. Especially the MSEs of both 2-par models being close to 0 is a very promising sign for their predictive ability. Secondly, the wet octanol models are slightly better than the dry octanol models. The difference here is less pronounced and might not be statistically significant, but the RMSE and the slope are both improved for the wet octanol model, which may indicate that it is indeed better suited to model the experimental conditions of the partition coefficient determination.

These results now have to be validated on the external test set comprised of SAMPL6 compounds. While less chemically diverse, this set of molecules has not been used for training of the corrections, and it is important to analyse the performance of the models on these compounds.

6.3.3 SAMPL6.2 partition coefficients

Turning to the compounds of the SAMPL6 challenge, the partition coefficients could be experimentally measured only for 11 of the 24 compounds. By calculating the Gibbs energies of each tautomer and conformation of each compound in the respective solvents and applying a partition function approach the log *P* values could be calculated according to equation (60). In the original SAMPL5 challenge an artificial gas phase reorganization term was used to calculate the Gibbs energies of transfer by conducting gas phase optimizations of the solute structures for the respective solvents and determining the resulting structures' gas phase energies. Conversely, here the Gibbs energy of transfer was directly calculated from the Gibbs energies of the solutes in solution because the gas phase energies should cancel exactly. The calculated partition coefficients for the SAMPL6.2 compounds in comparison with the experimental results are shown in Fig. 21. The individual results for each compound applying the different models are shown in Table 16 while the statistical metrics are summarized in Table 15.

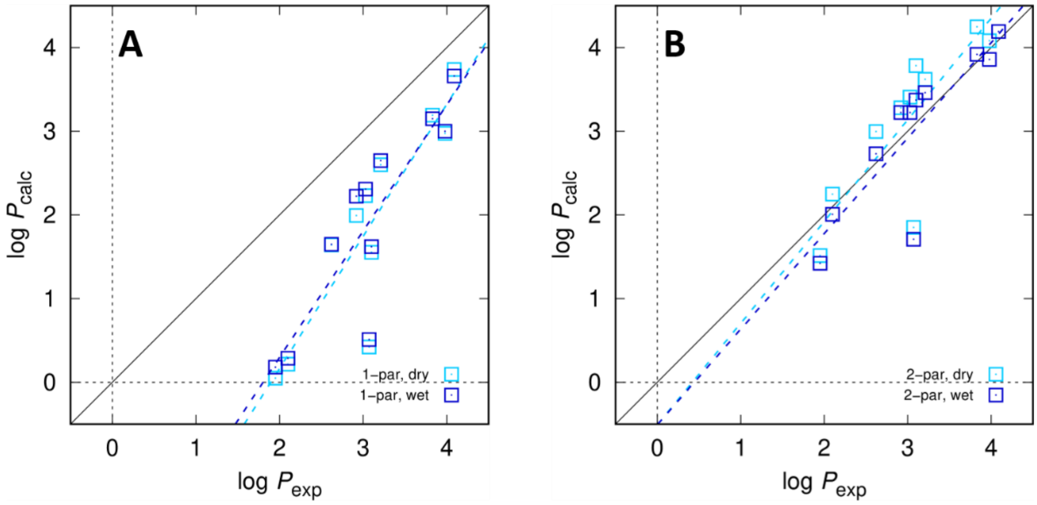


Fig. 21: Partition coefficients log *P* between water and octanol calculated using EC-RISM vs. the experimental values of the SAMPL6.2 set of compounds for both the 1-par (A) and the 2-par correction (B). Data generated using the dry octanol model is shown in light blue while those generated using the wet octanol model is shown in dark blue. Raw data are provided as part of OR_07. Optimized solution phase structures are the same as those used in the original SAMPL6 challenge gathered in OR_05. Figure adapted from [5].

Table 15: Statistical metrics for log *P* predictions (root-mean-square error RMSE, mean absolute error MAE, mean signed error MSE, slope *m*', intercept *b'*, and coefficient of determination R^2 from descriptive regression) for the four different models. Table adapted from [5].

Model	RMSE	MAE	MSE	*m*'	*b*'	R^2
dry, 1-par	1.38	1.21	-1.21	1.58	-2.99	0.79
wet, 1-par	1.32	1.15	-1.15	1.51	-2.72	0.77
dry, 2-par	0.54	0.45	0.15	1.22	-0.51	0.73
wet, 2-par	0.47	0.31	-0.07	1.14	-0.51	0.73

For both the 1-par model and the 2- par model there is only a very small difference between the results generated with the dry and the wet octanol models. There is, however, a massive difference between the 1- par and the 2- par model results. For the 1- par model there is a clear deviation between the experimental and the calculated partition coefficients that increases further with increasing hydrophilicity, as measured through the slope of 1.58 and 1.51 for the dry and the wet model, respectively. While the RMSE is still only slightly above 1.3, these models appear to underestimate the interaction of octanol with more polar solvents.

Table 16: Individual experimental and corresponding predicted log *P* values for all models. Material from: 'N Tielker, D Tomazic, L Eberlein, S Güssregen, SM Kast, The SAMPL6 challenge on predicting octanol–water partition coefficients from EC-RISM theory, J Comput-Aided Mol Des, published 2020, Springer' [5].

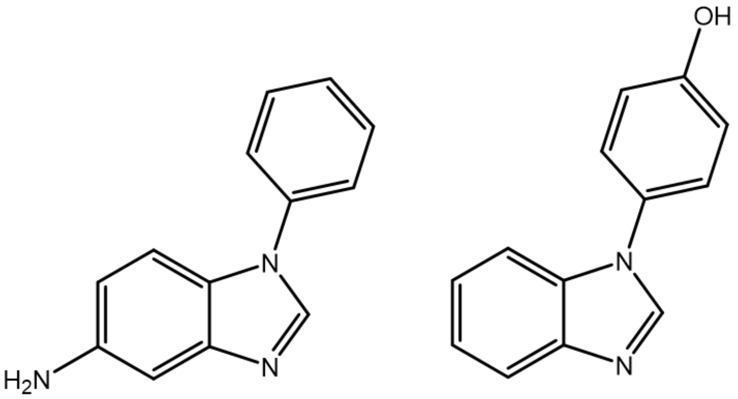
	log P_{exp}	dry, 1-par	wet, 1-par	dry, 2-par	wet, 2-par
SM02	4.09	3.74	3.66	4.56	4.19
SM04	3.98	2.97	3.00	4.08	3.86
SM07	3.21	2.60	2.65	3.62	3.46
SM08	3.10	1.55	1.62	3.78	3.37
SM09	3.03	2.23	2.31	3.41	3.22
SM11	2.10	0.22	0.29	2.25	2.01
SM12	3.83	3.19	3.15	4.25	3.92
SM13	2.92	1.99	2.22	3.28	3.22
SM14	1.95	0.05	0.18	1.51	1.42
SM15	3.07	0.42	0.51	1.85	1.71

SM16	2.62	1.64	1.65	3.00	2.73

The 2- par models on the other hand have much better agreement between experimental and predicted values for all compounds in the SAMPL6 log *P* challenge. Between the two models all statistical metrics are nearly the same, but the wet octanol model has a small edge across the board. With an RMSE of 0.47 pK units, this model places second among all physics-based models submitted in the SAMPL6 challenge [156]. Additionally, the low MSE and the slope close to unity are evidence for a very balanced model.

For this set of data it is more difficult to conceive a suitable null hypothesis, because there is no physically justifiable guess for an "average" log *P*. However, considering that the challenge was about drug-like molecules, a reasonable guess might be the mean partition coefficient of a library of drug-like compounds. Ghose et. al give an average log *P* of 2.52 for the Comprehensive Medicinal Chemistry database [157]. A challenge participant using no further resources and simply guessing that every compound has that partition coefficient would have achieved a respectable 32nd place out of 93 total submissions with an RMSE of 0.87. This value lies right between the RMSEs of the 1- par and the 2- par models, further supporting the assumption that two parameters are necessary for the octanol model.

One unfortunate outlier, the compound SM15, is responsible for the model not performing even better in the challenge. This outlier is surprising for another reason: the compound SM15 is structurally very similar to the compound SM14, which, while slightly underpredicted as well, was predicted significantly more accurately. Both structures are depicted in Fig. 22.



SM14 **SM15**

Fig. 22: Chemical structures of the outlier compound SM15 and the structurally related compound SM14.

The fact that this compound is the only alcohol in the SAMPL6 log *P* data set made reinvestigation of the performance on the alcohols contained in the MNSOL Gibbs energy of transfer data set advisable. The predicted partition coefficients of all SAMPL6.2 compounds and the alcohols contained in the MNSOL database, as well as the individual errors in the Gibbs energies of solvation for each MNSOL alcohol are shown in Fig. 23.

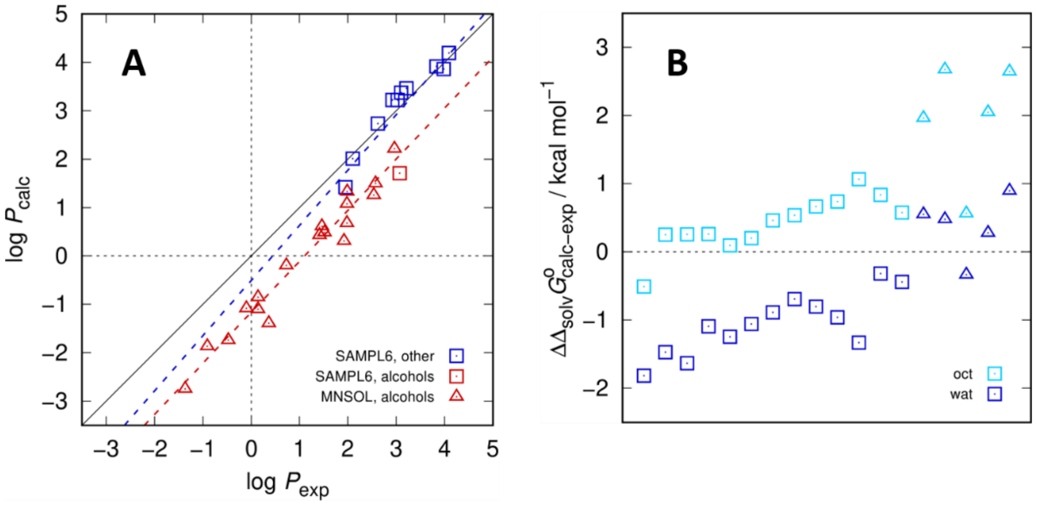


Fig. 23: Partition coefficients log *P* between water and octanol calculated using EC-RISM vs. the experimental values of the SAMPL6.2 set of compounds (squares) and the alcohols in the MNSOL database (triangles) for the best performing wet 2-par model (A). Alcohols are depicted in red while all other compound classes are depicted using blue. Also shown are the individual errors in the Gibbs energy of solvation in the two respective solvents for the MNSOL compounds (B). Aliphatic alcohols are depicted using squares while aromatic alcohols are depicted using triangles.

The MNSOL contains a number of aliphatic and aromatic alcohols, but regardless of the type or the absolute log *P* of the alcohol in question, the deviation between predicted and experimental value is almost constant. Regression of all alcoholic compounds in both sets yields a slope of 1.03, effectively unity, but a deviation of -1.16 pK units. In addition, since for the MNSOL compounds the individual experimental Gibbs energies of solvation are available there was a chance that the reason for this nearly constant offset could be determined by looking at the errors in the two solvents. It appears, however, that while the difference in the errors remains similar for every compound, leading to similar errors in the eventual log *P*, the Gibbs energy of solvation is at times predicted better in octanol, at times in water, and at times even with a similar magnitude in the error but of opposite sign. Especially surprising is the error in the related compounds *i*-, *o*-, and *m*-cresol. While the experimental Gibbs energies of solvation are very close, deviating by only 0.6 kcal·mol^{-1}, the predicted values are reasonable in water, but in octanol vary by as much as 1.8 kcal·mol^{-1}.

A larger database of alcoholic compounds might be able to shed some more light on this issue, especially since there appears to be a significant difference between aliphatic and aromatic alcohols. In the former the error in the Gibbs energy of solvation in octanol is larger, and negative, while in the latter the error in the Gibbs energy of solvation in water is larger and positive. For the immediate future, the reparameterization of the GAFF force field that is used throughout this work should be of higher priority and with respect to the difference between aliphatic and aromatic alcohols the addition of new atom types may have to be considered as well. The additional molecules recently published by the SAMPL6 challenge authors, many of which are alcohols, but also other compounds with a greater dynamic log *P* range than that investigated in this challenge may prove helpful in this endeavor [158,159].

The good performance of the other SAMPL6 compounds compared to the performance on the MNSOL Gibbs energies of transfer can be explained by the fact that only three scaffolds account for nine of the eleven compounds in the SAMPL6 data set. This should not detract from the good performance of the best-performing EC-RISM model, because obtaining the correct trends in a series of closely related compounds is a challenge itself. But the applicability to other compound classes should be investigated further in the future.

A second task that was not officially part of the SAMPL6 log *P* challenge was the determination of the relative tautomer stabilities in the different solvents. While no experimental values were determined for this property, there was an interest in how different methods would predict the tautomer stabilities in the polar and non-polar solvents. EC-RISM had already shown reasonable results in the SAMPL2 challenge on tautomer stabilities, although more recent results using high-level coupled cluster gas phase energies show a mixed picture [6,43]. Nevertheless, for this set of compounds the relative tautomer energies could be calculated by applying the partition function only to the conformations and not to the tautomers as well. The relative free energies of the different tautomers considered for each compound compared to the energetically most favorable one are collected in Table 17.

Table 17: Calculated Gibbs free energies of the neutral microstates relative to the most favorable tautomer (microstate) of each compound for both solvents (in kcal mol^{-1}). In contrast to the calculation of the partition coefficients where special treatment is not necessary, we here made sure that individual conformations undergoing a protonation shift during QC optimization were manually assigned to the correct microstate before evaluation of the partition function.

Microstate	Water	Octanol (wet, 2-par)	Octanol (dry, 2-par)	Octanol (wet, 1-par)	Octanol (dry, 1-par)
SM02_micro002	0.00	0.00	0.00	0.00	0.00
SM02_micro003	5.16	5.57	5.66	5.65	5.71
SM02_micro007	6.18	8.86	8.80	10.30	10.40
SM04_micro003	0.00	0.00	0.00	0.00	0.00
SM04_micro004	8.45	9.81	9.74	10.68	10.76
SM04_micro009	11.10	11.72	11.78	12.15	12.24
SM07_micro002	8.97	10.59	10.61	11.63	11.78
SM07_micro003	6.75	7.97	8.00	8.34	8.41
SM07_micro004	0.00	0.00	0.00	0.00	0.00
SM08_micro008	10.26	24.63	24.61	32.59	33.52
SM08_micro010	5.69	6.05	6.56	4.70	4.89
SM08_micro011	0.00	0.00	0.00	0.00	0.00
SM09_micro002	6.79	9.55	9.45	11.45	11.57

SM09_micro003	0.00	0.00	0.00	0.00	0.00
SM09_micro011	5.60	6.02	6.09	6.46	6.55
SM11_micro005	0.00	0.00	0.00	0.00	0.00
SM11_micro028	7.14	8.07	8.21	8.46	8.61
SM11_micro029	14.81	17.69	17.68	18.81	18.93
SM11_micro030	26.91	34.04	34.12	36.10	36.40
SM12_micro002	4.73	5.21	5.32	5.35	5.43
SM12_micro011	5.76	8.48	8.42	10.04	10.14
SM12_micro012	0.00	0.00	0.00	0.00	0.00
SM13_micro005	0.00	0.00	0.00	0.00	0.00
SM13_micro007	6.23	6.28	6.31	6.69	6.76
SM13_micro009	8.01	10.72	10.51	12.78	12.84
SM14_micro001	0.00	0.00	0.00	0.00	0.00
SM14_micro005	28.76	37.41	37.02	41.99	42.23
SM15_micro001	9.24	19.80	18.80	26.68	26.76
SM15_micro002	0.00	0.00	0.00	0.00	0.00
SM16_micro002	0.00	0.00	0.00	0.00	0.00
SM16_micro003	12.41	13.39	13.61	12.68	12.79
SM16_micro007	6.75	11.48	11.49	13.61	13.93

The results of the tautomer stability analysis show that there is no change in the lowest energy microstate, regardless of the solvent. There is however a difference in the relative energies of the energetically higher tautomers. It does not lead to a flip, but for example for the compound SM16 an energy difference of 1.91 kcal·mol^{-1} between the second and the third lowest tautomers, micro003 and micro007, in the best performing octanol model turns into a 5.66 kcal·mol^{-1} difference in water. Other tautomer pairs converge, for example SM08's two higher ranked tautomers are predicted to have an energy difference of 18.58 kcal·mol^{-1} in octanol, an energy difference that would make the energetically higher ranked microstate completely irrelevant, but in water this difference shrinks to "only" 4.57 kcal·mol^{-1}.

At least for the compounds investigated in the SAMPL6 challenge there does not appear to be a shift of the most energetically favored microstate between the two phases. For other compounds, especially ones that only have a low energy difference between the tautomer forms, this can change. For certain compounds there is even experimental evidence that the tautomer found in water is not the same as in an organic phase and knowledge of this is of vital importance to predict the true partition coefficient of a molecule [160,161].

6.4 SAMPL5 revisited

6.4.1 Introduction

The improvements that had been made over the course of the SAMPL6 challenges, especially with regards to the pK_a model and the treatment of the electrostatics made a reinvestigation of the distribution coefficients from the SAMPL5 challenge promising. While the cyclohexane as an apolar solvent does not benefit from the improved electrostatics, the water model and especially the pK_a model derived from it have been greatly improved since the time of the SAMPL5 challenge. Additionally, the computational resources available had increased significantly, making it possible to calculate distribution coefficients for all three batches with both the old SAMPL5 setup and the new and improved water and pK_a models [162].

6.4.2 Computational Details

To compare the predictive power of the old and the new models two different setups were used. In the following they will be referred to as the “SAMPL5 setup” and the “SAMPL6 setup”, respectively. For the SAMPL6 setup the MP2/6-311+G(d,p) water and pK_a models from the SAMPL6 challenge were used unchanged and all of the EC-RISM and 3D RISM settings employed for the SAMPL6 setup are identical to those used in the SAMPL6.2 challenge. However, a new cyclohexane PMV correction was trained where the vacuum conformations were explicitly reoptimized at the B3LYP/6-311+G(d,p) level of theory to account for molecular reorganization upon entering the solvent using Gaussian 09 Rev. A02 [142]. In the original SAMPL5 setup, the vacuum conformation was identical to the conformation in cyclohexane instead. Unlike in the original SAMPL5 challenge the entire set of compounds including batch 2 were investigated. For the SAMPL5 setup the most abundant tautomer states of batch 2 were conformationally sampled and the conformation with the lowest PCM energy treated with the same EC-RISM setup as described in chapter 4.2. In contrast, for the SAMPL6 setup every tautomer state generated using MoKa [134] was conformationally sampled using the workflow described there, and up to five conformations of each of the tautomers were taken into account using a partition function approach to account for the high conformational flexibility of the compounds included in the SAMPL5 challenge. Another important difference between the SAMPL5 and SAMPL6 setup is the calculation of the transfer free energies. In the original SAMPL5 setup these were calculated from solvation free energies, i.e. the free energy difference between the molecule in vacuum and in water and octanol, respectively. For the SAMPL6 setup they were instead calculated directly from the Gibbs energies of the molecules in their respective solvents. Formally this makes no difference, because the energies of the molecule in vacuum should cancel, however the vacuum conformations were originally generated separately from both the water and the cyclohexane conformations, leading to an artificial reorganization energy difference.

6.4.3 Results

In total four different cyclohexane models were trained using between a single and three free parameters to fully capture the range of possible corrections instead of only two models trained during the SAMPL5 challenge. These models are denoted as 1-par for the model using only a single parameter scaling the PMV (eq. (83)), 2-par (eq. (82)) and 2-par-I (eq. (76)) for the models with an additional parameter scaling the 3D RISM excess chemical potential or a linear offset, respectively, and 3-par for a model with all three parameters (eq. (75)). The 2-par-I and the 3-par models trained during the original challenge are denoted with a (5) after their respective identifier, but the general form of the correction is identical to those without a (5). The results for the Gibbs energies of solvation in cyclohexane using the different models are depicted in Fig. 24 while the parameter sets and the statistical metrics for each model are shown in Table 18.

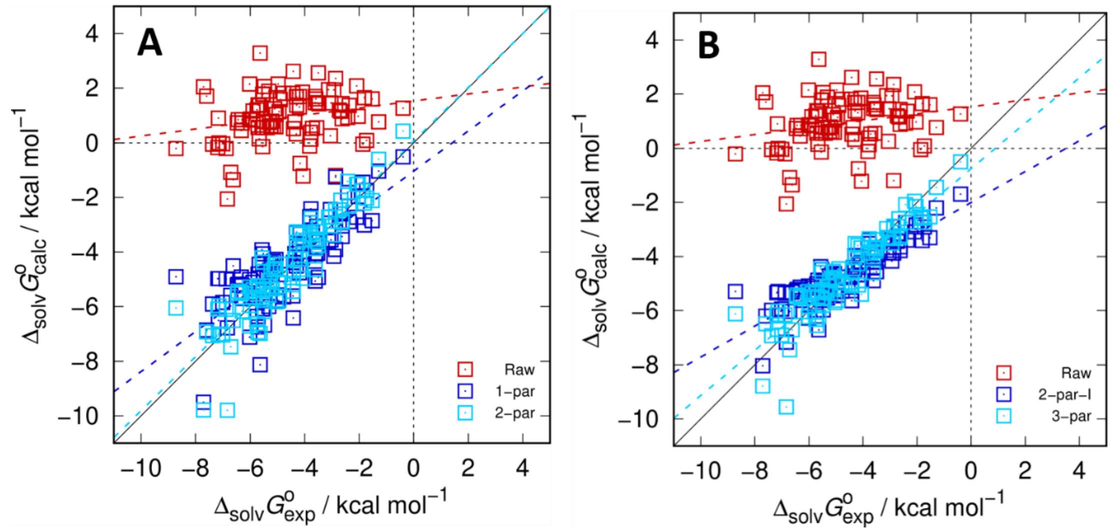


Fig. 24: Gibbs energies of solvation in cyclohexane calculated using EC-RISM vs. the experimental results from the MNSOL database. Uncorrected data is shown using red squares, panel A shows the corrections without an intercept parameter while panel B shows the corrections with one. Corrections using a parameter directly scaling the excess chemical potential are shown in light blue and corrections without such a parameter in dark blue. Raw and calculated data are provided as part of OR_09. Optimized solution and gas phase structures are provided as OR_10. Figure adapted from [6].

Table 18: Regression parameters of optimized EC-RISM-based Gibbs energy of solvation models (c_μ, c_V / kcal mol^{-1} Å^{-3}, c_q / kcal mol^{-1} e^{-1}) along with statistical metrics (root-mean-square error RMSE / kcal mol^{-1}, mean absolute error MAE / kcal mol^{-1}, mean signed error MSE / kcal mol^{-1}, slope *m*', intercept *b*' / kcal mol^{-1}, and coefficient of determination R^2 from descriptive regression). c_V corresponds to PMVs computed using the total correlation function with an experimental isothermal compressibility of 1.197·10^{-9} Pa^{-1} for cyclohexane and the RISM estimate of 0.717062·10^{-9} Pa^{-1} for water. Material from: 'N Tielker, L Eberlein, G Hessler, KF Schmidt, S Güssregen, SM Kast, Quantum–mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges?, J Comput-Aided Mol Des, published 2020, Springer' [6].

Solvent	RMSE	MAE	MSE	*m*'	*b*'	R^2	c_μ	c_V	c_q	c_d
Water										

All	2.04	1.43	-0.26	1.00	-0.35	1.00	-	-0.10251	-15.728	-
Neutrals	1.56	1.13	-0.36	0.97	-0.47	0.89	-	-	-	-
Anions	3.07	2.46	0.01	1.10	7.18	0.94	-	-	-	-
Cations	2.98	2.10	0.02	0.96	-2.62	0.85	-	-	-	-
Cyclohexane										
Uncorrected	5.86	5.60	5.60	0.13	1.53	0.05	-	-	-	-
1-par	1.07	0.86	0.20	0.73	-1.04	0.62	-	-0.14923	-	-
2-par	0.77	0.58	0.11	0.99	0.06	0.83	2.0184	-0.17795	-	-
2-par-I	0.90	0.73	0.00	0.57	-2.00	0.76	-	-0.10894	-	-1.6593
2-par-I(5)	0.88	0.70	0.00	0.59	-1.94	0.77	-	-0.10811	-	-1.6566
3-par	0.68	0.50	0.00	0.84	-0.75	0.83	1.8516	-0.14692	-	-1.0842
3-par(5)	0.76	0.56	0.00	0.84	-0.73	0.84	1.8444	-0.14703	-	-1.0479

For the models that were already used in the original SAMPL5 challenge the results obtained for the PMV correction for cyclohexane are very similar [1]. This had to be expected because the improved electrostatics in the 3D RISM iterations do not have an effect on the calculations for this apolar solvent and only the addition of the gas-phase optimized conformations slightly improves the results for the 3-par model. One notable addition is the 2-par model. This model had not been investigated during the original SAMPL5 challenge but had performed very well when used with another apolar solvent, octanol, during the SAMPL6 challenge [5]. For the compounds in the MNSOL database this model also performs best for cyclohexane with a low RMSE, a slope of 0.99 and an intercept of only 0.06, which are the best statistical metrics among all models and point towards a very balanced model. It should be noted that even the 1-par model's performance, as judged by its RMSE, is significantly better than the best-performing 2-par model for octanol. This may, however, simply be caused by the different compounds for which experimental data is available in the MNSOL database. Generally improved performance for the log $D_{7.4}$ predictions could be expected for this model. The partition and distribution coefficients calculated using both the old models originally considered in chapter 4 and the new models developed in this chapter are depicted in Fig. 25. The corresponding statistical metrics are shown in Table 19.

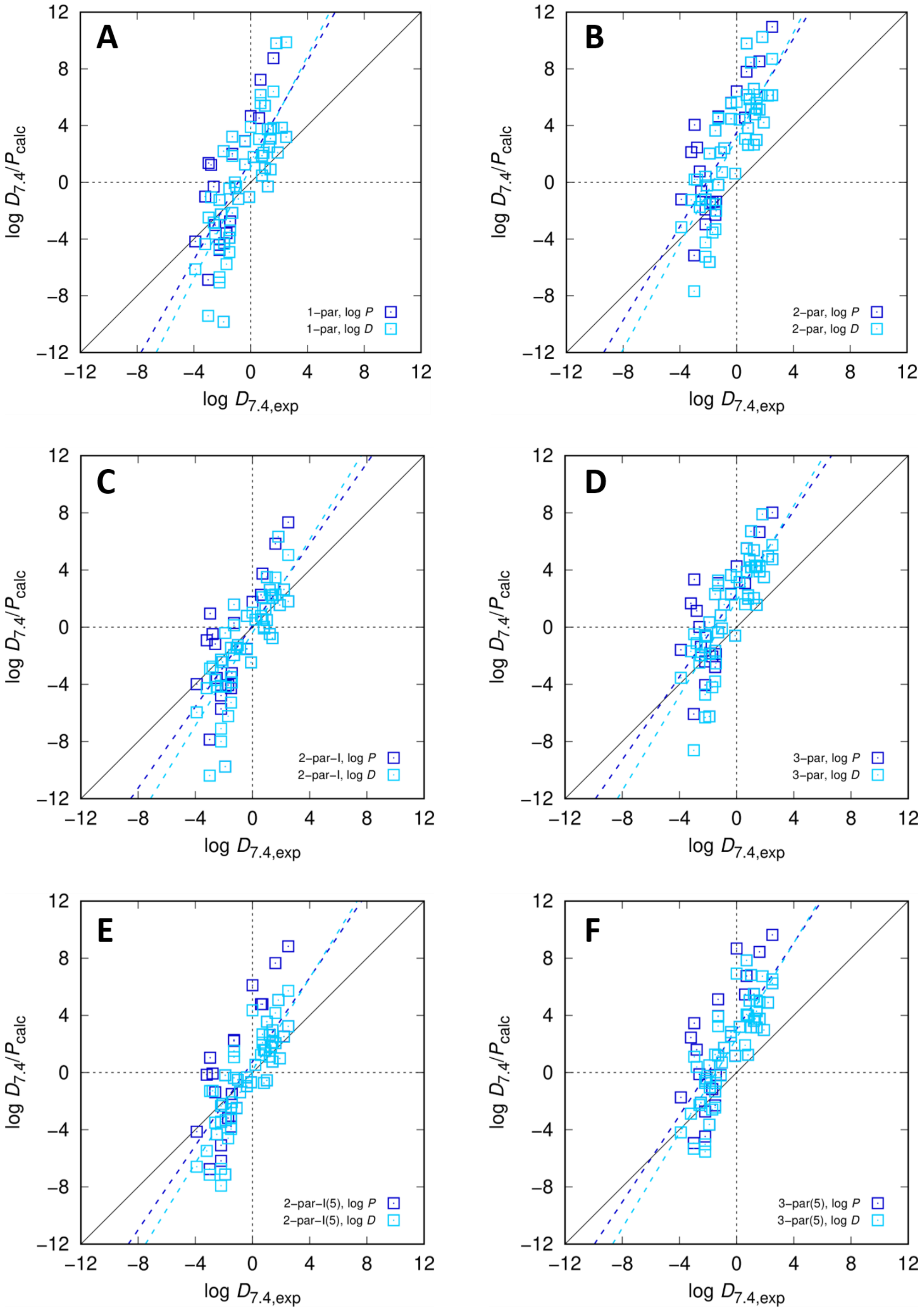


Fig. 25: Partition (dark blue) and distribution coefficients (light blue) calculated using EC-RISM compared with the experimental results of the SAMPL5 challenge. Panel A shows results generated using the 1-par correction, panel B the 2-par results, panel C the 2-par-I and panel D the 3-par results. Raw data are provided as part of OR_09. Newly generated optimized solution phase structures are provided as OR_10. Figure adapted from [6].

Table 19: Statistical metrics (root-mean-square error RMSE, mean absolute error MAE, mean signed error MSE, and slope m, intercept b, and coefficient of determination R^2 from descriptive regression) for the SAMPL5 challenge results for all compounds. Material from: 'N Tielker, L Eberlein, G Hessler, KF Schmidt, S Güssregen, SM

Kast, Quantum–mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges?, J Comput-Aided Mol Des, published 2020, Springer' [6].

Setup	Observable	Cyclohexane mod.	Batches	RMSE	MSE	MAE	R^2	*m*'	*b*'
SAMPL6	log *P*	1-par	0+1	2.29	0.13	1.77	0.63	1.56	0.43
			2	4.74	3.18	4.01	0.54	2.04	3.56
		2-par	0+1	3.18	2.37	2.59	0.66	1.52	2.64
			2	5.87	5.14	5.53	0.59	1.82	5.44
		2-par-I	0+1	1.99	-0.65	1.57	0.62	1.31	-0.49
			2	2.83	1.10	2.09	0.53	1.58	1.31
		3-par	0+1	2.44	1.49	1.97	0.63	1.36	1.68
			2	4.15	3.47	3.93	0.60	1.56	3.67
	log $D_{7.4}$	1-par	0+1	2.45	-0.59	1.88	0.77	1.89	-0.12
			2	4.26	2.62	3.58	0.63	2.18	3.05
		2-par	0+1	2.88	1.65	2.49	0.74	1.85	2.09
			2	5.36	4.59	4.98	0.66	1.95	4.94
		2-par-I	0+1	2.44	-1.37	1.73	0.74	1.64	-1.04
			2	2.48	0.55	1.66	0.64	1.71	0.80
		3-par	0+1	2.33	0.77	1.91	0.72	1.69	1.13
			2	3.65	2.92	3.38	0.68	1.69	3.17
SAMPL5	log *P*	2-par-I(5)	0+1	1.99	-0.09	1.48	0.61	1.35	0.09
			2	2.83	1.67	2.32	0.52	1.39	1.81
		3-par(5)	0+1	2.86	2.08	2.41	0.65	1.41	2.30
			2	3.86	2.98	3.54	0.67	1.81	3.27
	log $D_{7.4}$	2-par-I(5)	0+1[a]	2.25	-0.86	1.63	0.71	1.60	-0.54
			2	2.44	0.48	1.99	0.69	1.81	0.77
		3-par(5)	0+1[b]	2.59	1.31	2.29	0.70	1.66	1.67
			2	4.68	4.17	4.29	0.56	1.40	4.32

[a-b] Corrected results for SAMPL5 setup, original values [1] for RMSE, MSE, R^2, *m*', *b*':
[a] 2.15, -0.53, 0.59, 1.36, -0.34;
[b] 2.76, 1.64, 0.59, 1.42, 1.87.

There are a number of key takeaways from these predictions. The results obtained for the batch 2 compounds are generally worse than those for the smaller batch 0 and batch 1 compounds, regardless of the model applied. While for the original SAMPL5 setup the lack of different conformers and tautomers might have been an explanation for this finding, the results are the same for the new SAMPL6 setup that explicitly accounts for this. Additionally, for both setups there is no clear improvement when taking the molecules' pK_a into account. While the RMSE is lowered for most models there are always individual compounds that are predicted significantly worse, leading to much higher slopes of the regression. Additionally, given the model errors inherent to the water and cyclohexane models, there is no statistically significant difference between the RMSEs obtained using the SAMPL5 and the SAMPL6 setup even though the PMV corrections and the pK_a model were improved since the original SAMPL5 challenge. Most surprisingly, the 2-par model which had been expected to perform well judging from the statistical metrics of the training set is even worse than the simple 1-par ansatz for both the log *P* and log *D* prediction. The predicted and experimental results for each individual compound are shown in Table 20.

Table 20: Experimental distribution coefficients and calculated partition and distribution coefficients for all models for which both SAMPL5 and SAMPL6 data exists. Table adapted from [6].

SAMPL5 ID	log $D_{7.4,\text{exp}}$	log P 2-par-I(5)	log P 3-par(5)	log P 1-par-	log P 2-par	log P 2-par-I	log P 3-par	log $D_{7.4}$ 2-par-I(5)	log $D_{7.4}$ 3-par(5)	log $D_{7.4}$ 1-par	log $D_{7.4}$ 2-par	log $D_{7.4}$ 2-par-I	log $D_{7.4}$ 3-par
Batch 0													
003	1.90	1.17	3.19	2.09	4.22	1.54	3.51	1.17	3.19	2.09	4.22	1.54	3.51
015	-2.20	-5.28	-2.87	-4.76	-1.92	-4.79	-2.41	-8.08	-5.67	-7.07	-4.23	-7.10	-4.72
017	2.50	3.39	6.39	3.20	6.14	1.81	4.75	3.39	6.39	3.20	6.14	1.81	4.75
020	1.60	1.98	3.83	3.83	5.12	2.28	3.91	1.98	3.83	3.83	5.12	2.28	3.91
037	-1.50	-3.79	-2.31	-3.91	-2.29	-4.27	-2.79	-3.95	-2.47	-4.92	-3.30	-5.27	-3.80
045	-2.10	-2.42	-0.64	-2.26	-0.22	-2.43	-0.67	-2.42	-0.64	-2.26	-0.22	-2.43	-0.67
055	-1.50	-3.13	-1.31	-3.91	-1.53	-3.50	-1.65	-3.13	-1.31	-3.91	-1.53	-3.50	-1.65
058	0.80	-0.83	1.16	0.47	2.64	0.03	2.00	-0.83	1.16	0.47	2.64	0.03	2.00
059	-1.30	-0.25	1.32	-2.17	-0.17	-1.96	-0.36	-0.25	1.32	-2.17	-0.17	-1.96	-0.36
061	-1.45	-1.19	0.08	-2.76	-1.37	-3.22	-1.89	-1.91	-0.65	-3.39	-2.00	-3.86	-2.53
068	1.40	0.95	3.33	0.91	2.99	-0.76	1.57	0.95	3.33	0.91	2.99	-0.76	1.57
070	1.60	7.32	8.25	8.76	8.52	5.84	6.65	3.56	4.48	6.40	6.16	3.48	4.29
080	-2.20	-3.42	-0.71	-4.69	-1.21	-4.11	-1.40	-3.42	-0.71	-4.69	-1.21	-4.11	-1.40
Batch 1													
004	2.20	2.60	4.96	3.85	6.12	2.64	4.96	2.60	4.96	3.84	6.12	2.63	4.95
005	-0.86	-1.44	1.68	-1.17	2.41	-1.54	1.58	-1.44	1.68	-1.18	2.41	-1.54	1.58
007	1.40	2.91	4.90	3.73	5.59	2.22	4.30	2.91	4.90	3.73	5.59	2.22	4.30
010[a]	-1.70	-3.45	-1.43	-3.60	-1.38	-4.05	-2.03	-5.88	-3.85	-5.77	-3.55	-6.23	-4.21
011[b]	-2.96	1.03	3.43	1.36	4.05	0.95	3.34	-1.67	0.74	-2.48	0.21	-2.89	-0.50
021	1.20	1.22	3.72	-0.28	2.65	-0.48	2.04	1.22	3.72	-0.28	2.65	-0.48	2.04
026[c]	-2.60	-2.08	-0.82	-0.31	0.77	-1.18	0.02	-5.02	-3.76	-2.82	-1.74	-3.69	-2.49
027	-1.87	-3.44	-1.16	-4.29	-1.48	-4.12	-1.83	-3.44	-1.16	-4.34	-1.53	-4.17	-1.88
042	-1.10	0.40	2.63	0.01	2.12	-1.44	0.83	0.40	2.63	0.01	2.12	-1.44	0.83
044	1.00	-0.74	2.97	1.00	5.21	0.50	4.19	-0.74	2.97	1.00	5.21	0.50	4.19
046	0.20	0.70	3.38	1.79	4.42	0.53	3.17	0.70	3.38	1.79	4.42	0.53	3.17
047	-0.40	-0.35	2.53	1.26	4.48	0.79	3.64	-0.35	2.53	1.26	4.48	0.79	3.64
048	0.90	1.47	5.07	2.08	5.86	1.28	4.74	1.47	5.07	2.08	5.86	1.28	4.74
056	-2.50	-1.10	1.12	-3.02	-0.63	-3.56	-1.37	-1.10	1.12	-3.63	-1.24	-4.17	-1.98
060[d]	-3.90	-4.19	-1.79	-4.17	-1.21	-3.99	-1.58	-6.86	-4.45	-6.13	-3.17	-5.95	-3.54
063	-3.00	-6.93	-5.06	-6.88	-5.15	-7.86	-6.08	-8.77	-6.90	-9.41	-7.68	-10.39	-8.61
071	-0.10	-0.99	1.02	-1.03	0.61	-2.47	-0.60	-1.02	0.99	-1.04	0.61	-2.48	-0.60
072	0.60	3.49	4.30	4.53	4.55	2.27	3.09	-0.05	0.76	3.04	3.07	0.78	1.60
081	-2.20	-6.02	-4.20	-4.41	-2.96	-5.72	-4.05	-7.69	-5.86	-6.68	-5.23	-7.99	-6.32
090	0.80	2.04	4.46	1.87	3.82	-0.08	2.23	2.04	4.46	1.87	3.82	-0.08	2.23
Batch 2													
002	1.40	2.17	4.35	3.07	5.22	2.06	4.21	2.17	4.35	3.07	5.22	2.06	4.21
006	-1.02	0.20	1.41	-0.28	0.71	-1.26	-0.09	0.20	1.41	-0.28	0.71	-1.26	-0.09
013	-1.50	-2.53	1.28	-0.44	3.64	-1.45	2.31	-2.53	1.28	-0.44	3.64	-1.45	2.31
019	1.20	2.81	5.61	3.74	6.59	2.61	5.38	2.77	5.57	3.74	6.59	2.61	5.38
024	1.00	3.46	6.75	5.40	8.43	3.51	6.70	3.46	6.75	5.40	8.43	3.51	6.70
033	1.80	5.06	6.72	9.80	10.24	6.33	7.90	5.06	6.72	9.80	10.24	6.33	7.90
049	1.30	1.80	3.81	2.50	4.79	2.25	4.25	1.80	3.81	2.50	4.79	2.25	4.25
050	-3.20	-0.11	2.49	-1.00	2.12	-0.91	1.67	-5.58	-2.98	-4.36	-1.24	-4.27	-1.69
065	0.70	1.88	7.06	6.16	9.79	0.54	5.53	1.88	7.06	6.16	9.79	0.54	5.53
067	-1.30	1.40	3.15	3.23	4.54	1.59	3.26	0.17	1.94	3.23	4.54	1.59	3.26
069	-1.30	2.34	5.18	2.01	4.64	0.28	3.08	0.95	3.79	1.86	4.49	0.13	2.93
074	-1.90	-6.61	-3.04	-9.85	-5.62	-9.76	-6.25	-6.61	-3.04	-9.85	-5.62	-9.76	-6.26
075	-2.80	1.35	3.07	1.22	2.46	-0.48	1.15	-0.36	1.37	-1.05	0.18	-2.75	-1.13
082	2.50	8.17	9.06	12.15	10.96	7.34	8.02	4.94	5.84	9.88	8.69	5.06	5.75
083[e]	-1.90	-	-	-	-	-	-	-	-	-	-	-	-
084	0.00	3.79	6.52	4.66	6.42	1.77	4.25	1.25	3.97	3.90	5.67	1.02	3.50

SAMPL5 ID	log $D_{7.4,exp}$	log P 2-par-I(5)	log P 3-par(5)	log P 1-par-	log P 2-par	log P 2-par-I	log P 3-par	log $D_{7.4}$ 2-par-I(5)	log $D_{7.4}$ 3-par(5)	log $D_{7.4}$ 1-par	log $D_{7.4}$ 2-par	log $D_{7.4}$ 2-par-I	log $D_{7.4}$ 3-par
085	-2.20	-2.33	-0.57	-1.24	0.39	-2.29	-0.56	-8.14	-6.39	-1.24	0.39	-2.29	-0.56
086	0.70	4.15	6.59	7.23	7.80	3.74	5.52	2.89	5.32	5.58	6.15	2.09	3.87
088	-1.90	-1.46	-0.62	2.19	2.02	-0.41	0.35	-1.46	-0.62	2.19	2.02	-0.41	0.35
092	-0.40	-0.71	3.52	2.91	5.61	-1.51	2.33	-0.71	3.52	2.87	5.56	-1.55	2.28

These surprising findings must be adequately explained. The reason for the deviation of the regression slopes is the inverse sigmoidal-shaped distribution of the predicted values. While the experimental results indicate a dynamic range from -3.90 to 2.50 the predictions cover a much larger range in both directions. This is further amplified by the directional nature of the distribution coefficient because accounting for the ionic species can only shift the log *D* towards more negative values. Molecules that are already underpredicted in the log *P* thus become even more underpredicted, an effect that cannot be offset by the improved predictions of previously overestimated log *P* values in the RMSEs. At the edges of the dynamic range the errors exceed 4 p*K* units, leading to bad statistical metrics and strongly deviating regression slopes.

When looking at the SAMPL5 and the SAMPL6 pK_a models used there are some changes in the predictions here, too. Unfortunately, no experimental pK_a values exist for these compounds, but comparing the predictions of both setups shows them to have good correlation, while those of the former are on average lower by about 1.15 pK units. Comparing these results with predictions from a different source, in this case using pK_a values empirically predicted using Chemicalize [163] shows that both methods have reasonable agreement with the empirical predictions with RMSEs of 2.10 and 2.07, respectively. The acidity constants calculated using EC-RISM with the SAMPL5 and SAMPL6 model as well as those predicted by Chemicalize are depicted in Fig. 26. The statistical metrics for the two different EC-RISM based models are shown in Table 21.

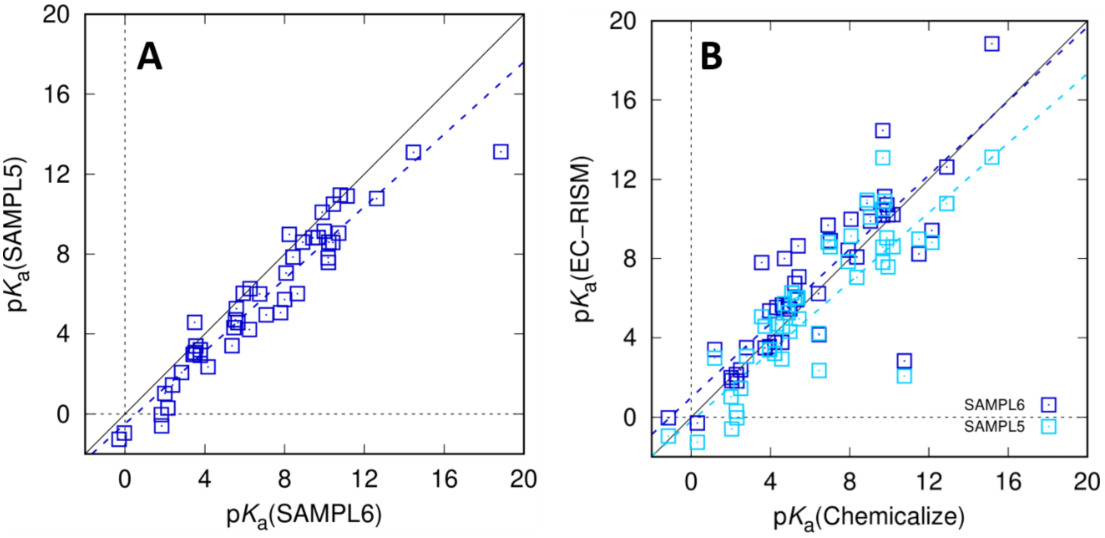


Fig. 26: (A) shows acidity constants calculated with the SAMPL5 setup compared with those calculated with the SAMPL6 setup. (B) shows the results obtained with the SAMPL5 setup (light blue) and the SAMPL6 setup (dark blue) compared with those predicted with the empirical Chemicalize tool. Raw data are provided as part of OR_09. Optimized structures are provided as OR_10. Figure adapted from [6].

Table 21: Statistical metrics (root-mean-square error RMSE, mean absolute error MAE, mean signed error MSE, and slope *m*, intercept *b*, and coefficient of determination R^2 from descriptive regression) for the SAMPL5 challenge results for all compounds.

pK_a model	RMSE	MSE	MAE	R^2	m'	b'
SAMPL5	2.07	-0.57	1.54	0.72	0.88	0.21
SAMPL6	2.10	0.58	1.45	0.73	0.94	0.99

While the RMSEs of 2.07 and 2.10 of the SAMPL5 and the SAMPL6 setup, respectively, might suggest that the two pK_a models are nearly identical, this is not actually the case. The MSEs reveal that the SAMPL5 acidity constants tend to underestimate the reference values obtained using Chemicalize while the SAMPL5 acidity constants tend to overestimate them by approximately the same absolute value. The higher predicted pK_a values of the SAMPL6 setup lead to different effects for acids and bases: acids will be predicted to have a lower fraction of the ionic species at a pH of 7.4 so their log *D* is closer to the log *P*, while bases will be predicted to have a higher fraction of the ionic species and their log *D* will be reduced more compared to the log *P*. Since there are 33 basic and only 14 acidic predicted pK_a values this leads to a stronger effect of the pK_a on the already slightly lower partition coefficients predicted using the SAMPL6 setup when calculating the distribution coefficients. While this difference is part of the reason for the different results for the SAMPL5 and the SAMPL6 setup, this cannot explain the large differences between the predicted values from both setups and the experimental results.

When discussing deviations between experimental and theoretical results it is also necessary to consider deviations in the theoretical modeling of the real experiment that might cause some errors. The cyclohexane model used here is a pure organic phase while the experimentally determinable water fraction in cyclohexane is estimated between $3.20 \cdot 10^{-4}$ and $3.75 \cdot 10^{-4}$ [164]. While this might be assumed to be a negligibly small water content, for the more polar compounds this might have a significant effect because there is some evidence that the water molecules may form complexes with them, improving their solubility in the organic phase [51]. Klamt *et al.* who submitted the best-performing model during the original SAMPL5 challenge studied the effect of these small water concentrations but found only some very minor improvements for some of their predicted distribution coefficients and an improved RMSE of 2.08 from 2.11 before accounting for the water content of the organic phase [165].

Unless otherwise mentioned, in the following only the best-performing model of the SAMPL6 setup, 2-par-I, is discussed. Comparing the predicted distributions coefficients obtained with EC-RISM with those of the Klamt *et al.* the agreement is significantly better than with the experimental data. While there is an offset towards lower predicted values for EC-RISM, with an RMSE of 1.77 the agreement between the two models is very reasonable even for the most hydrophilic and lipophilic compounds that have the worst agreement with the experimental data. The good agreement between the two theoretical approaches and the lack of improvement upon consideration of the water content in the cyclohexane phase for the approach used by Klamt *et al.* show that there is no indication of a systematic deficiency of the apolar phase model used in this work. The distribution coefficients calculated using EC-RISM and those predicted by Klamt *et al.* are depicted in Fig. 27.

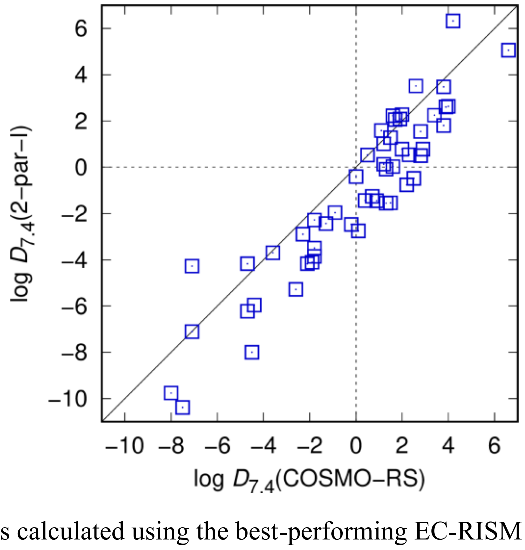


Fig. 27: Distribution coefficients calculated using the best-performing EC-RISM model compared with predictions from the best-performing model overall during the original SAMPL5 challenge [165]. Material from: 'N Tielker, L Eberlein, G Hessler, KF Schmidt, S Güssregen, SM Kast, Quantum–mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges?, J Comput-Aided Mol Des, published 2020, Springer' [6].

The agreement between both approaches and the experimental data is significantly stronger when ignoring the outliers. Klamt *et al.* investigated a set that was generated by excluding the eight worst outliers, yielding an RMSE of only 1.57. For EC-RISM by removing the seven worst outliers, SAMPL5_033, 010, 015, 037, 063, 074, and 081, the RMSE drops to merely 1.37 and the MSE to 0.12. For both models these outliers are clearly the main cause for the large errors in the predictions. These results indicate that while for the medium range of distribution coefficients the theoretical models are performing well, for very hydrophilic and very lipophilic side there is a large discrepancy between the model predictions and the experimental results. Effects such as dimerization are not included in either computational model and the formation of such species can influence the partitioning behavior in a way that would not be seen in the predictions. On the experimental side some more general potential issues with the experimental design have been discussed in the literature, such as a low equilibration time and the possibility of detector saturation [52,166]. A more fundamental problem has been mentioned by Hill and Young when they compared computationally predicted and experimentally measured octanol-water distribution coefficients [167]. They find a very similar distribution of the predicted values with high accuracy in the mid-range and large errors going towards the extremes. The reason might be a combination of low solubility in one phase with a high solubility in the other phase leading to larger uncertainties of the measurement for the former, and non-ideal solution behavior in the latter phase. The reason for the large errors for these compounds might thus at least partially be found in the experimental measurements.

Finally, it is useful to give a null hypothesis for the prediction of the distribution coefficients to assess the overall performance of the computational predictions. Unfortunately, for the log $D_{7.4}$ no precise average value could be found, but the analysis of 18,428 compounds by Hill and Young suggests a similar average distribution coefficient as found for the partition coefficients [167]. Using the same value of 2.5 as a null hypothesis gives a considerably large RMSE of 3.45 which would not have been enough to break into the top half of the submissions. However, this truly blind null hypothesis does not take into account the composition of the SAMPL5 set of compounds. A *post hoc* analysis would have to conclude that the distribution of distribution coefficients in the SAMPL5 challenge significantly differs from the one in the octanol-water dataset investigated by Hill and Young. Using an adjusted null hypothesis of exactly zero for the “guessed” distribution coefficient results in an RMSE of 1.78, easily winning the challenge. While the model RMSEs after removal of the worst outliers is lower than this, it is difficult to draw conclusions from this fact since, by removing the worst performing compounds only the compounds with low contributions to the RMSE remain. This shows that while for simpler problems good solutions have already been found, there is still much work to do to bring experimental and predicted distribution coefficients into accordance with each other.

7 Partial Molar Volume correction for molecules under extreme conditions

7.1 Introduction

The successful application of the PMV correction for a wide range of different solvents and the ability to calculate not just Gibbs energies of solvation, the property the models were originally trained for, but any kind of property that can be derived from the Gibbs energy of the molecule in solution made it attractive to investigate further areas where this method could be applied. EC-RISM had been used for some time to investigate the properties of molecules under high hydrostatic pressure [88,168,169]. The behavior of biomolecules under high pressure is not just of academic interest, enabling researchers to probe protein dynamics and folding [69], but may also be biologically relevant when discussing the potential origin of life in the deep sea or extraterrestrial biology [170]. Even today complex life exists in the deep sea or even deeper in the oceanic crust and has adapted to the higher pressures present there [171,172].

As long as only relative properties are investigated there is no need for a PMV correction, because of the error cancellation if the PMV does not change, but to make EC-RISM as versatile under these conditions as it is at atmospheric pressure there was some need to develop such a model. Even before this work it did not seem likely that the 1 bar correction would hold up under high hydrostatic pressure because of its physical origin: The error in the excess chemical potential is related to the overestimation of the energy of cavity formation and this in turn is directly related to the pressure (or density) of the solvent.

The most obvious problem with the development of a PMV correction for non-atmospheric pressures is the lack of experimental reference data. The MNSOL database contains only Gibbs energies of solvation under normal conditions and an investigation of the available literature did not yield any usable data either. Even if some reference data at higher pressure had been found, EC-RISM is used for calculations of systems with a pressure of up to 10 kbar, a pressure that cannot be reached experimentally. Since experimental data was unavailable, a way to generate reference data using a method that is inherently able to include pressure information was needed. The method chosen for this purpose were TI reference calculations as it is well suited to this task for a number of reasons. TI calculations are routinely used to calculate Gibbs energies of solvation with high accuracy and it is possible to minimize model error by using the same force field (GAFF) and water model (SPC/E) that is used in the 3D RISM calculations. Furthermore, while it is impossible to prove due to the lack of available experimental data, there is no evidence that TI calculations suffer from significant issues at higher pressures.

However, using TI reference energies gives rise to a different problem: the PMV correction is already trained using experimental data at a pressure of 1 bar. It is in theory possible to calculate the TI reference energies for the entire MNSOL database but doing so would be both prohibitively expensive as well as not ideal compared to using experimental reference data for the 1 bar correction. For this reason, the choice was made to use the experimental 1 bar reference data to train a 1 bar PMV correction for EC-RISM and then fit the high-pressure correction in such a way that the energy difference between the 1 bar TI calculations and the PMV-corrected 1 bar EC-RISM calculations ideally remains constant over the entire pressure range. This can be achieved by generating high pressure reference data from TI calculations and adding the 1 bar energy difference between the EC-RISM and TI results. The reference data is the calculated as

$$\Delta_{\mathrm{solv}} G^{0}_{\mathrm{ref}}(p) = \Delta_{\mathrm{solv}} G^{0}_{\mathrm{TI}}(p) + (\Delta_{\mathrm{solv}} G^{0}_{\text{EC-RISM, 1 bar}}(1\ \mathrm{bar}) - \Delta_{\mathrm{solv}} G^{0}_{\mathrm{TI}}(1\ \mathrm{bar}))\ . \tag{82}$$

The correction then has the form of

$$\Delta_{\mathrm{solv}} G^{0}_{\text{EC-RISM,hp}}(p) = \Delta_{\mathrm{solv}} G^{0}_{\text{EC-RISM,1 bar}}(p) + c_{\mathrm{hp}} \left(p - 1\ \mathrm{bar}\right) V_m(p), \tag{83}$$

where the subscript “hp” denotes the high-pressure PMV correction while the subscript “1 bar” denotes the usual PMV correction under normal conditions. This expression recovers the results of the original PMV correction unchanged at a pressure of 1 bar and scales the results with respect to the pressure at higher than atmospheric pressures. The target function

$$\{c_{\mathrm{hp}}\} = \arg\min \left\lfloor \sum_{\mathrm{molecules}} \left(\Delta_{\mathrm{solv}} G^{0}_{\mathrm{ref}}(p) - \Delta_{\mathrm{solv}} G^{0}_{\text{EC-RISM,1 bar}}(p) - c_{\mathrm{hp}} \left(p - 1\ \mathrm{bar}\right) V_m(p)\right)^2 \right\rfloor \tag{84}$$

then yields the high-pressure parameter.

7.2 Computational details

To calculate a representative sample of the neutral molecules in the Minnesota Solvation Database, 51 neutral molecules from different substance classes were manually chosen for reference TI calculations.

To set up the simulations, 4167 SPC/E water molecules were placed in a box with an edge length of 50 Å around a single molecule of the solute, the structure of which corresponded to the minimum PCM structure used for EC-RISM calculations, using packmol 1.1.2.023. For the solute molecules AM1-BCC charges and parameters from the General Amber Force Field (GAFF 1.7, i.e. equivalent to GAFF 1.4) were used [173]. All simulations were conducted using NAMD 2.11 [174]. During the simulations 1-4 interactions were scaled by 0.833333 and all water bonds were kept rigid using the SETTLE algorithm [175]. Lennard-Jones interactions were gradually switched off between 10 and 12 Å while the electrostatic interactions were accounted for using a 4th order Particle Mesh Ewald (PME) algorithm with a grid spacing of 1.0 Å [176]. The temperature was held constant at 298.15 K using Langevin dynamics while the pressure was set to the target pressure (1, 100, 500, 1000, 2000, 3000, 4000, 5000, 7500 or 10000 bar) using the Nosé-Hoover-Langevin piston pressure control [177,178].

The system was minimized for 5000 steps with the conjugate gradient and line search algorithm implemented in NAMD. The system was then equilibrated for 0.4 ns with a time step of 2 fs with all interactions between solute and solvent molecules fully switched off. During the TI simulations the coupling parameter was first increased linearly for the Lennard-Jones interactions between 0 and 1 in steps of 0.1. After these interactions were fully switched on the electrostatic interactions were gradually switched on using the same step size. Hysteresis was performed in the reverse order, first turning off the electrostatic and then the Lennard-Jones interactions. In each lambda window the system was first equilibrated for 60 ps before being simulated for another 0.4 ns. The resulting data was analyzed using a bootstrapping scheme [179].

The EC-RISM calculations were conducted using the settings described in chapter 5.2, but using the solvent susceptibilities developed by T. Pongratz with the isothermal compressibilities calculated according to eq (39), and densities and dielectric constants taken from the work of Floriano and Nascimento [69,180]. The values used for each pressure are tabulated in Table 1 of the work of Pongratz *et al.* [4].

7.3 Results

First the results for the Gibbs energies of solvation from PMV-corrected EC-RISM calculations and TI calculations at normal pressure should be compared, because if the EC-RISM predictions are significantly better than the TI results the basis for using TI as a reference for higher pressures is called into question. The comparison of the Gibbs energies of solvation calculated from the two different approaches is shown in Fig. 28 and the individual Gibbs energies of solvation including the statistical uncertainty from the TI calculations in Table 22.

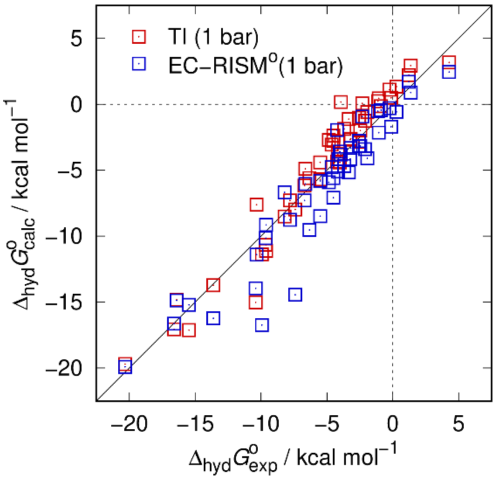


Fig. 28: Gibbs free energies of hydration at 1 bar calculated with EC-RISM and TI in comparison with experimental data taken from the Minnesota Solvation Database. Raw data for all EC-RISM and TI calculations in this chapter are provided in OR_11. Optimized structures are provided as OR_12. Figure adapted from [4].

Table 22: Individual experimental and corresponding predicted Gibbs energies of hydration in units of kcal·mol-^{1} calculated using TI and EC-RISM for a pressure of 1 bar, as well as statistical uncertainties ΔTI for the Gibbs energies of hydration generated from TI. Calculated TI Gibbs energies of hydration for all pressures are provided in as OR10.3 in OR-10.

MNSOL	Compound	Exp.	TI	ΔTI	EC-RISM
0018cyc	cyclohexane	1.23	2.23	0.15	1.71
0025buta	1-butene	1.38	2.95	0.13	0.91
0034hex	1-hexyne	0.29	1.34	0.16	-0.56
0036tol	toluene	-0.89	-0.14	0.19	-0.43
0042ant	anthracene	-4.23	-4.34	0.25	-1.94
0047pro	1-propanol	-4.83	-2.68	0.17	-5.90
0050met	*t*-butanol	-4.51	-2.77	0.18	-5.58
0053phe	Phenol	-6.62	-4.88	0.19	-6.04
0056mcr	*m*-cresol	-5.49	-4.38	0.19	-5.79
0060dim	dimethylether	-1.92	-0.59	0.19	-4.10
0068ani	anisole	-2.45	-1.99	0.17	-3.14
0072but	butanal	-3.18	-2.61	0.17	-3.30
0074ben	benzaldehyde	-4.02	-4.40	0.18	-3.59
0086eth	acetic acid	-6.70	-6.16	0.16	-7.28
0098met	methyl pentanoate	-2.57	-2.85	0.20	-3.18
0103eth	ethylamine	-4.50	-2.38	0.13	-7.07
0107tri	trimethylamine	-3.23	-2.68	0.14	-4.14
0109pip	piperazine	-7.40	-7.96	0.18	-14.43
0115dip	dipropylamine	-3.66	-1.86	0.19	-4.68

0118ani	aniline	-5.49	-5.69	0.19	-8.47
0119met	2-methyl pyridine	-4.63	-3.05	0.18	-4.76
0131nit	1-nitropropane	-3.34	-1.11	0.14	-5.16
0134nit	nitrobenzen	-4.12	-2.85	0.17	-4.14
0138pro	1-propanethiol	-1.05	0.47	0.16	-2.13
0139thi	thiophenol	-2.55	-1.01	0.19	-2.77
0154dif	1,1-difluoroethane	-0.11	0.50	0.12	-1.69
0168chl	2-chloropropane	-0.25	1.12	0.14	-0.29
0174chl	chlorobenzene	-1.12	0.25	0.16	-0.53
0187dib	dibromomethane	-2.30	0.07	0.21	-0.92
0211tri	1,1,1-trifluoropropan-2-ol	-4.16	-3.78	0.16	-5.09
0213bis	bis(2-chloroethyl) sulfide	-3.92	0.19	0.20	-3.80
0217wat	water	-6.31	-5.59	0.12	-9.51
0401amia	1,1-dimethyl-3-phenylurea	-9.63	-10.65	0.24	-10.13
0402adn	9-methyladenine	-13.60	-13.71	0.24	-16.23
0403thi	1-methylthymine	-10.40	-15.01	0.21	-13.95
0406oct	octafluoropropane	4.28	3.18	0.14	2.45
n015	3-aminoaniline	-9.92	-11.37	0.19	-16.73
n191	uracil	-16.59	-17.05	0.19	-16.63
n201	5-trifluoromethyluracil	-15.46	-17.12	0.24	-15.21
test1003	butylnitrate	-2.10	-1.25	0.18	-3.40
test1007	alachlor	-8.20	-8.51	0.29	-6.67
test1048	propanil	-7.80	-7.30	0.24	-8.76
test1049	pyrazon	-16.40	-14.83	0.27	-14.86
test1051	sulfometuron methyl	-20.30	-19.68	0.33	-19.89
test2025	phthalimide	-9.61	-11.09	0.20	-9.15
test3007	2-methoxybezoic acid	-10.32	-7.61	0.19	-11.4

The results show that the TI predictions are of similar quality as the EC-RISM data. While there is a shift in the sign of the error, EC-RISM tends to predict lower energies while TI calculations tend to predict higher energies, this does not matter because the correction will not be trained directly on the TI reference data but using the reference data generated as described above. The Gibbs energies of solvation at high pressures calculated using EC-RISM both with and without the additional high-pressure correction are depicted in Fig. 29. Table 23 shows the statistical metrics for the different approaches and the high-pressure correction.

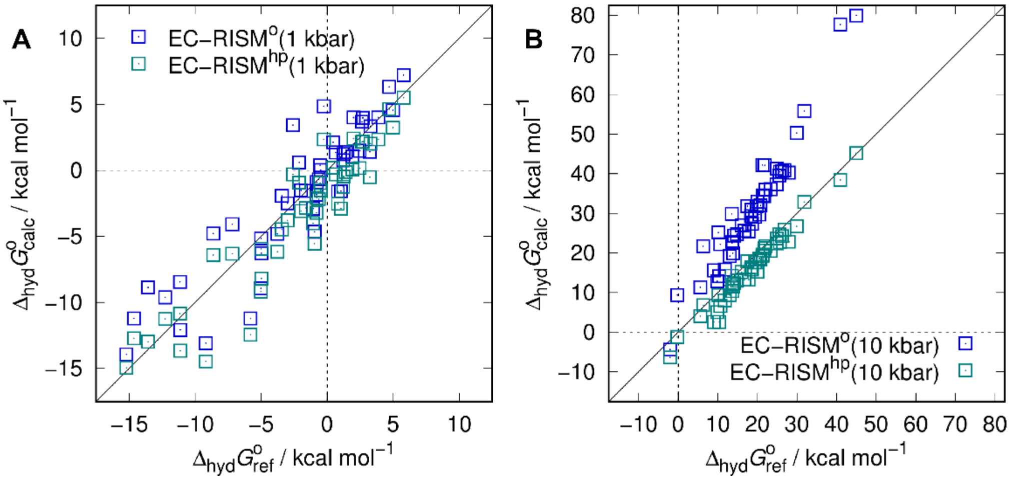


Fig. 29: Gibbs free energy of solvation calculated with EC-RISM in comparison with TI reference calculations at pressures of 1 kbar (A) and 10 kbar (B). Figure reprinted from [4] (https://doi.org/10.1016/j.bpc.2019.106258).

Table 23: Statistical metrics for Gibbs free energies of hydration calculated by EC-RISM, where the superscript “0” indicates the 1 bar and the superscript “hp” indicates the *p*-dependent PMV correction, and by TI in comparison with experimental values for 1 bar and for Gibbs free energies of solvation calculated by EC-RISM in comparison with TI results (root mean square error RMSE, mean absolute error MAE, mean signed error MSE, slope *m*’, intercept *b*’, and coefficient of determination R^2 from descriptive regression) for PMV-corrected and high pressure-corrected results. Corrected EC-RISM Gibbs energies of hydration are provided as OR10.4 in OR-10.

Model	RMSE	MAE	MSE	*m*'	*b*’	*R*^{2}
1 bar (exp. ref.)						
TI	1.56	1.24	0.65	0.87	-1.28	0.94
EC-RISM^0	1.98	1.32	-0.86	0.88	0.12	0.89
1 bar (TI ref.)						
EC-RISM^0	2.44	1.99	1.52	0.90	-1.99	0.89
EC-RISM^{hp}	2.44	1.99	1.52	0.90	-1.99	0.89
1 kbar (TI ref.)						
EC-RISM^0	2.39	1.79	-0.36	0.91	0.15	0.82
EC-RISM^{hp}	2.34	1.85	1.26	0.93	-1.41	0.86
10 kbar (TI ref.)						
EC-RISM^0	14.34	12.64	-12.64	1.62	0.97	0.92
EC-RISM^{hp}	2.92	2.41	2.30	1.05	-3.25	0.97

After training of the correction, the Gibbs energies of solvation predicted from EC-RISM are in line with the Gibbs energies of solvation from the TI reference calculations. While there is a small increase in the RMSE from 2.44 kcal·mol^{-1} at a pressure of 1 bar to 2.92 kcal·mol^{-1} at a pressure of 10 kbar using the TI as a reference, the error obtained when using only the 1 bar correction is much higher at 14.34 kcal·mol^{-1}. The increase in the error at very high pressures might be due to the increased polarization of the molecules in the EC-RISM calculations whereas in the TI reference calculations the same AM1-BCC charges were used for every pressure. Another potential reason is the fact that the pressures used in the correction are not spaced evenly but instead there are as many pressures between 1 and 2000 bar as there are between 2000 and 10000 bar.

The lack of experimental data at high hydrostatic pressures makes not only the development of the PMV correction for high pressures difficult, but also the validation. However, the autoprotolysis constant of water is known to shift to lower values under increasing pressure [181]. P. Kibies was able to show that application of the high pressure PMV correction is able to correct both the trend and the magnitude of the shift for different force fields under investigation when compared to the 1 bar PMV correction [4,87]. This is strong evidence for the necessity of the high pressure PMV correction and its correct training with the TI reference data. More evidence for the high pressure PMV correction should be forthcoming as it is used in more research areas in the future.

.

8 SUMMARY AND CONCLUSION

In the course of this work a number of improvements to the scope of applications and the accuracy of the EC-RISM solvation model were made, many of which can and already have been applied to other fields of research in the group. These improvements fall into one of the following three categories:

1. Development of a workflow to generate solvent susceptibilities for diverse solvents and training and testing of a PMV correction for it.
2. Development and testing of models to generate accurate acidity constants and other physicochemical properties for compounds with an arbitrary number of conformational, tautomer, and ionic states.
3. Development of a general method to gain access to reference data for environmental conditions for which little or no experimental data is available.

The methods described in chapter 3 can be used as default methods to generate solvent susceptibilities for many different solvents if the usual 1D-RISM-based workflow fails, or to generate alternative solvent models. In combination with a well-performing water model, the prediction of partition coefficients makes these nonaqueous solvents more accessible, as they can otherwise be difficult to investigate due to lack of experimental data. In the future, further investigation into methods to develop PMV corrections for solvents without sufficient training data is also necessary, e.g. to handle solvent mixtures of varying composition or a wider range of nonaqueous solvents.

Building on the PMV corrections developed by D. Tomazic, different combinations of level of theory, basis set, and PMV correction model were examined. Using the PMV correction enables the calculation of accurate Gibbs energies for small, organic compounds in solution, which can in turn be used to calculate molecular properties such as partition coefficients. To calculate acidity constants and distribution coefficients a second model was developed that accounts for the proton's Gibbs energy of hydration, that cannot be accessed with EC-RISM. The active participation in the SAMPL series of challenges, where such properties have to be predicted, made it possible to take advantage of the opportunities provided by the blind challenges for model development and improvement. The access to curated experimental data sets specifically developed for prediction by theoretical methods should be used for further developments in the future. Especially the comparison of predictions made using EC-RISM with those of other methods can shed light on the strengths and weaknesses of EC-RISM, and potential avenues of improvement, as has been done in this work.

One thing that the reanalysis of the SAMPL5 dataset in chapter 6.4 showed, is that there is no clear improvement in the quality of the predictions, despite the advances in the EC-RISM methodology, PMV and pK_a models, and the treatment of multiple conformational or tautomeric states. The question in how far the employed theoretical models mimic the experimental reality, where e.g. non-ideal mixture effects may play a role, needs to be considered in future works. Moreover, it might be useful to also repeat experiments when new experimental equipment has been developed to minimize uncertainty in the reference data needed to train and validate theoretical models. To advance the prediction and understanding of physicochemical properties, close collaboration of computational and experimental groups is necessary.

Use of the PMV correction is not limited to the properties investigated in this work. As the Gibbs energy is a fundamental thermodynamic property any molecular property that can be derived from them can, in principle, now be calculated using EC-RISM. Furthermore, while spectroscopic properties calculated from the wave function of the molecules in solution are not directly affected by the PMV correction, the conformational and tautomer populations of the molecules under investigation are.

The high-pressure correction developed in this work makes it possible to calculate the properties of molecules under high hydrostatic pressure, conditions under which living organisms must adapt but can still exist. Research on topics such as base pairing stability under high hydrostatic pressure has already been carried out using this new correction [182]. Furthermore, the new correction gave the initial spark to the development of a temperature-dependent partial molar volume correction that is still under development but will make it possible to calculate the properties of molecules under higher temperatures and their enthalpies.

Finally, some of the “results” obtained over the course of this dissertation are not just the development and improvement of certain models and approaches to calculating physicochemical properties, but also methods to handle large amounts of data, the standardized generation of conformations of flexible molecules for use with EC-RISM and the script-based analysis of the results obtained from these calculations. Applying these advances to the three categories PMV correction development, physicochemical property prediction, and biophysical molecules under extreme conditions in the future will increase the speed, scope, and accuracy of EC-RISM even further.

9 REFERENCES

[1] Tielker N, Tomazic D, Heil J, Kloss T, Ehrhart S, Güssregen S, Schmidt KF, Kast SM (2016) J Comput Aid Mol Des 30:1035-1044.

[2] Tielker N, Eberlein L, Güssregen S, Kast SM (2018) J Comput Aid Mol Des 32:1151-1163.

[3] Tielker N, Eberlein L, Chodun C, Güssregen S, Kast SM (2019) J. Mol. Model. 25:139.

[4] Pongratz T, Kibies P, Eberlein L, Tielker N, Hölzl C, Imoto S, Beck Erlach M, Kurrmann S, Schummel PH, Hofmann M, Reiser O, Winter R, Kremer W, Kalbitzer HR, Marx D, Horinek D, Kast SM (2020) Biophys Chem. 257:106258.

[5] Tielker N, Tomazic D, Eberlein L, Güssregen S, Kast SM (2020) J Comput Aid Mol Des 34:453-461.

[6] Tielker N, Eberlein L, Hessler G, Schmidt KF, Güssregen S, Kast SM J. Comput Aid Mol Des, https://doi.org/10.1007/s10822-020-00347-5

[7] Leo A, Hansch C, Elkins D (1971) Chem Rev 71:525-616.

[8] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Ad Drug Deliver Rev 23:3-25.

[9] Liu X, Testa B, Fahr A (2011) Pharm Res 28:962-977.

[10] Kloss T, Heil J, Kast SM (2008) J Phys Chem B 112:4337-4343.

[11] Beglov D, Roux B (1997) J Phys Chem B 101:7821-7826.

[12] Kovalenko A, Hirata F (1998) Chem Phys Lett 290:237-244.

[13] Sato H (2013) Phys Chem Chem Phys 15:7450-7465.

[14] Ratkova EL, Palmer DS, Fedorov MV (2015) Chem Rev 115:6312-6356.

[15] Sergiievskyi V, Jeanmairet G, Levesque M, Borgis D (2015) J Chem Phys 143:184116.

[16] Misin M, Fedorov MV, Palmer DS (2016) J Phys Chem B 120:975-983.

[17] Marenich AV, Kelly CP, Thompson JD, Hawkins GD, Chambers CC, Giesen DK, Winget P, Cramer CJ, Truhlar DG (2012) Minnesoate Solvation Database – version 2012, University of Minnesota, Minneapolis.

[18] Paul SM, Mytelka DS, Dunwiddie CT, Personger CC, Munos BH, Lindborg SR, Schacht AL (2010) Nat Rev Drug Discov 9:203-214.

[19] Morgan P, Van der Graaf PH (2012) Drug Discov Today 17:419-424.

[20] Singh SS (2006) Curr Drug Metab 7:165-182.

[21] Wenzel J, Matter H, Schmidt F (2019) J Chem Inf Model 59:1253-1268.

[22] Overington JP, Al-Lazikani B, Hopkins AL (2006) 5:993-996.

[23] Waterbeemd H, Smith DA, Beaumont K, Walker DK (2001) 44:1313-1332.

[24] Leeson PD, Springthorpe B (2007) Nat Rev Drug Discov 6:881-890.

[25] Price DA, Blagg J, Jones L, Greene N, Wager T (2009) Expert Opin Drug Metab Toxicol 5:921-931.

[26] Peters JU, Schnider P, Mattei P, Kansy M 2009 ChemMedChem 4:680-686.

[27] Gleeson MP (2007) J Med Chem 50:101-112.

[28] Young RJ, Green DV, Luscombe CN, Hill AP (2011) Drug Discov Today 16:822-830.

[29] Perozzo R, Folkers G, Scapozza L (2004) J Recept Signal Transduct Res 24:1-52.

[30] Waterbeemd H, Gifford E (2003) Nat Rev Drug Discovery 2:192-204.

[31] Wang Y, Xing J, Xu Y, Zhou N, Peng J, Xiong Z, Liu X, Luo X, Luo C, Chen K, Zheng M, Jiang H (2015) Q Rev Biophys 48:488-515.

[32] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V (2015) J Chem Inf Model 55:263-274.

[33] Hessler G, Baringhaus KH (2018) Molecules 23:2520.

[34] Davies JT (1950) J Phys Chem 54:185-204.

[35] Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN (2013) J Am Chem Soc 135:7296-7303.

[36] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2012) Adv Drug Deliv Rev 64:4-17.

[37] Cossi M, Barone V, Cammi R, Tomasi J (1996) Chem Phys Lett 255:327-335.

[38] Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) J Phys Chem A 101:3005-3014.

[39] Kühne TD (2014) Wiley Interdiscip Rev-Comput Mol Sci 4:391-406.

[40] Cramer CJ, Truhlar DG (1999) Chem Rev 99:2161-2200.

[41] Bashford D, Case DA (2000) Ann Rev Phys Chem 51:129-152.

[42] Zhou R, Berne BJ (2002) PNAS 99:12777-12782.

[43] Kast SM, Heil J, Güssregen S, Schmidt KF (2010) J Comput Aided Mol Des 24:343-353.

[44] Tomazic D (2016) PhD thesis. https://eldorado.tu-dortmund.de/handle/2003/34883

[45] https://samplchallenges.github.io/. Accessed 2020/10/19.

[46] Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) J Med Chem 51:769-779.

[47] Guthrie JP (2009) J Phys Chem B 113:4501-4507.

[48] Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) J Comput-Aid Molec Des 24:259-279.

[49] Geballe MT, Guthrie JP (2012) J Comput-Aid Molec Des 26:489-496.

[50] Guthrie JP (2014) J Comput-Aid Molec Des 28:151-168.

[51] Bannan CC, Burley KH, Chiu M, Shirts MR, Gilson MK, Mobley DL (2016) J Comput-Aid Molec Des 30:927-944.

[52] Rustenburg AS, Dancer J, Lin B, Feng JF, Ortwine DF, Mobley DL, Chodera JD (2016) J Comput Aided Mol Des 30:945-958.

[53] Grebe SKG, Singh RJ (2011) Clin Biochem Rev 32:5-31.

[54] Isik M, Levorse D, Rustenburg AS, Ndukwe IE, Wang H, Wang X, Reibarkh M, Martin GE, Makarov AA, Mobley DL, Rhodes T, Chodera JD (2018)) J Comput-Aid Molec Des 32:1117-1138.

[55] Isik M, Bergazin TD, Fox T, Rizzi A, Chodera JD, Mobley DL (2020) J Comput-Aid Molec Des 34:335-370.

[56] https://github.com/samplchallenges/SAMPL7 (last accessed 2020/10/19).

[57] Fanelli D (2012) Scientometrics 90:891-904.

[58] Hohenberg P, Kohn W (1964) Phys Rev 136:864-871.

[59] Hansen JP, Mc Donald IR (2007) Theory of simple liquids, 3rd Edition, Elsevier Academic Press, Amsterdam; Boston.

[60] Evans R (1979) Adv Phys 28:143-200.

[61] Gray C, Gubbins KE (1984) Theory of molecular fluids. Oxford University Press, New York, United States.

[62] Hirata F (2003) Molecular Theory of Solvation. Springer, Berlin, Germany.

[63] Workman H, Fixman M (1973) J Chem Phys, 58:5024-5030.

[64] Morita T, Hiroike K (1961) Prog Theor Phys 25:537-578.

[65] Chandler D, Andersen HC (1972) J Chem Phys 57:1930-1937.

[66] Lowden LJ, Chandler D (1974) J Chem Phys 61:5228-5241.

[67] Kast SM, Schmidt KF, Schilling B (2003) Chem Phys Lett 367:398-404.

[68] Du Q, Beglov D, Roux B (2000) J Chem Phys 104:796-805.

[69] Hölzl C, Kibies P, Imoto S, Frach R, Suladze S, Winter R, Marx D, Horinek D, Kast SM (2016) J Chem Phys 144:144104.

[70] Chandler D, McCoy JD, Singer SJ (1986) J Chem Phys 85:5977-5982.

[71] Cortis CM, Rossky PJ, Friesner J (1997) J Chem Phys 107:6400-6414.

[72] Kovalenko A, Hirata F (1999) J Chem Phys 110:10095-10112.

[73] Kovalenko A, Hirata F (1999) J Phys Chem B 103:7942-

[74] Kast SM, Kloss T (2008) J Chem Phys 129:236101.

[75] Singer S, Chandler D (1986) Mol Phys 55:621-625.

[76] Imai T, Kinoshita M, Hirata F (2000) J Chem Phys 112:9469-9478.

[77] Imai T (2007) Cond Matt Phys 10:343-361.

[78] Kirkwood JG, Buff FP (1951) J Chem Phys 19:774-777.

[79] Kusalik PG, Patey GN (1988) J Chem Phys 89:5843-5851.

[80] Chong SH, Hirata F (1997) J Phys Chem B 101:3209-3220.

[81] Imai T, Nomura H, Kinoshita M, Hirata F (2002) J Chem Phys B 106:7308-7314.

[82] Yamaguchi T, Matsuoka T, Koda S (2003) J Chem Phys 119:4437-4448.

[83] Heil J, Kast SM, (2015) J Chem Phys 142:114107.

[84] Chirlian LE, Francl MM (1987) J Comput Chem 8:894-905.

[85] Breneman CM, Wiberg KB (1990) J Comp Chem 11:361-373.

[86] Hoffgard F, Heil J, Kast SM (2013) J Chem Theory Comput 9:4718-4726.

[87] Patrick Kibies PhD thesis. https://eldorado.tu-dortmund.de/handle/2003/38208

[88] Frach R, Kast SM (2014) J Phys Chem A, 118:11620-11628.

[89] Frach R, Kibies P, Böttcher S, Pongratz T, Strohfeldt S, Kurrmann S, Koehler J, Hofmann M, Kremer W, Kalbitzer HR, Reiser O, Horinek D, Kast SM (2016) Angew Chem Int Ed 55:8757-8760.

[90] Imoto S, Kibies P, Rosin C, Winter R, Kast SM, Marx D (2016) Angew Chemie Int Ed 55:9534-9538.

[91] Güssregen S, Matter H, Hessler G, Lionta E, Heil J, Kast SM (2017) J Chem Inf Model 57:1907-1922.

[92] Palmer D, Frolov A, Ratkova E, Fedorov M (2010) J Phys Condens Matter 22:492101.

[93] Palmer D, Frolov A, Ratkova E, Fedorov M (2011) Mol Pharmaceutics 8:1423-1429.

[94] Truchon JF, Pettit BM, Labute P (2014) J Chem Theory Comput 10:934-941.

[95] Ng KC (1974) J Chem Phys 61:2680-2689.

[96] Sergiievskyi V, Jeanmairet G, Levesque M, Borgis D (2015) J Chem Phys 143:184116.

[97] Kathmann SM, Kuo IFW, Mundy CJ, Shenter GK (2011) J Phys Chem B 115:4369-4377.

[98] Kastenholz MA, Hünenberger PH (2006) J Chem Phys 124:224501.

[99] Lin YL, Aleksandrov A, Simonson T, Roux B (2014) J Chem Theory Comput 10:2690-2709.

[100] Beck TL (2013) Chem Phys Lett 561:1-13.

[101] Asthagiri D, Pratt LR, Ashbaugh HS (2003) J Chem Phys 119:2702-2708.

[102] Munte CE, Karl M, Kauter W, Eberlein L, Pham TV, Beck Erlach M, Kast SM, Kremer W, Kalbitzer HR (2019) Biophys Chem 254:106261.

[103] Tissandier MD, Cowen KA, Feng AY, Gundlach E, Cohen MH, Earhart AD, Coe JV (1998) J Phys Chem A 102:7787-7794.

[104] Zhang H, Jiang Y, Yan H, Cui Z, Chunhua Y (2017) J Chem Inf Model 57:2763-2775.

[105] Tielker N, Eberlein L, Chodun C, Güssregen S, Kast SM (2019) J Mol Model 25:139.

[106] Pracht P, Wilcken, R, Udvarhelyi, A, Rodde, S, Grimme, S (2018) J Comput-Aid Molec Des 32:1139–1149.

[107] Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) J Phys Chem A 107:9380-9386.

[108] Bochevarov AD, Watson MA, Greenwood JR (2016) J Chem Theory Comput 12:6001-6019.

[109] Heil J (2016) PhD dissertation. https://eldorado.tu-dortmund.de/handle/2003/35930.

[110] Scherrer RA, Howard SM (1977) J Med Chem 20:53-58.
[111] Kah M, Brown CD (2008) Chemosphere 71:1401-1408.
[112] Perilla JR, Goh BC, Cassidy CK, Liu B, Bernadi RC, Rudack T, Yu H, Wu Z, Schulten K (2015) Curr Opin Struct Biol 31:64-74.
[113] Coveney PV, Wan S (2016) Phys Chem Chem Phys 18:30236-30240.
[114] Ye X, Cui S, de Almeida VF, Khomami B (2013) J Mol Mod 19:1251-1258.
[115] Swope WC, Andersen HC, Berens PH, Wilson KR (1982) J Chem Phys 76:637-649.
[116] Miyamoto S, Kollman PA (1992) J Comp Chem 13:952-962.
[117] Berens PH, Mackay DHJ, White GM, Wilson KR (1983) J Chem Phys 79:2375-2389.
[118] Straatsma TB, Berendsen HJC (1988) J Chem Phys 89:5876-5886.
[119] Beutler TC, Mark AE, van Shaik RC, Gerber PR, van Gunsteren WF (1994) Chem Phys Lett 222:529-539.
[120] Zacharias M, Straatsma TP, McCammon JA (1994) J Chem Phys 100:9025-9031.
[121] Tielker N (2015) Master thesis, “Theoretische Modelle für Verteilungsgleichgewichte Kleiner Moleküle zwischen Wasser und *n*-Octanol”.
[122] Lang BE (2012) J Chem Eng Data 57:2221-2226.
[123] Martínez L, Andrade R, Birgin EG, Martínez JM (2009) J Comput Chem 30:2157-2164.
[124] Schuler LD, Daura X, van Gunsteren WF (2001) J Comput Chem 22:1205–1218.
[125] Aicart E, Tardajos G, Diaz Pena M (1981) J Chem Eng Data 26:22-26.
[126] Ryckaert JP, Ciccotti G, Berendsen HJC (1977) J Comput Phys 23:327-341.
[127] Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) J Chem Theory Comput 4:435-447
[128] Lide DR (2004) CRC handbook of chemistry and physics, 84th edition, CRC Press, Boca Raton, Florida.
[129] Dallos A, Liszi J (1995) J Chem Thermodyn 27:447.448.
[130] Lippold BC, Adel MS (1972) Arch Pharm 305:417-426.
[131] Talman JD (1978) J Comput Phys 29:35-48.
[132] Rossky PJ, Friedman HL (1980) J Chem Phys 72:5694-5700.
[133] Dierckx P (1982) SIAM J Numer Anal 19:1286-1304.
[134] Miletti F, Storchi L, Sforna G, Cruciani G (2007) J Chem Inf Model 47:2172-2181.
[135] Molecular Networks GmbH, Corina, https://www.mn-am.com/products/corina (last accessed 16/04/2020).
[136] RDKit: Open-source cheminformatics. http://www.rdkit.org (last accessed 03/02/2021).
[137] Ebejer J-P, Morris GM, Deane CM (2012) J Chem Inf Model 52:1146-1158.
[138] Sigalove G, Fenley A, Onufriev A (2006) J Chem Phys 124:124902.
[139] Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Hayik S, Roitberg A, Seabra G, Swails J, Götz AW, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wolf RM, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh MJ, Cui G, Roe DR, Mathews DH, Seeting MG, Salomon-Ferrer R, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, Kollman PA (2012) AMBER 12, University of Californai, San Francisco.
[140] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) J Comput Chem 25:1157-1174.
[141] Jakalian A, Jack DB, Bayly CI (2002) J Comput Chem 23:1623-1641.
[142] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Keith T, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG,

Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ, (2013) Gaussian 09 Rev. A.01, Wallingford CT.

[143] Klicić JJ, Friesner RA, Liu S-Y, Guida WC (2002) J Phys Chem 106:1327-1335.

[144] Maw S, Sato H, Ten-no S, Hirata F (1997) Chem Phys Lett 276:20-25.

[145] Sato H, Hirata F (1999) J Chem Phys 111:8548-8555.

[146] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Menucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Naktasuji 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 E, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham A, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, People JA (2004) Gaussian 03, Rev D.02, Gaussian Inc., Wallingford CT.

[147] Eckert F, Klamt A (2006) J Comput Chem 27:11-19.

[148] Small-Molecule Drug Discovery Suite 2017-2 (2017), Schrödinger LLC, New York.

[149] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone,V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09, Rev E.01. Gaussian, Inc., Wallingford CT.

[150] Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) J Phys Chem A 107:9380-9386.

[151] Chodun C (2018) Bachelor Thesis, „Optimierung von Modellen zur Vorhersage von pKa-Werten".

[152] Auch J (2018) Bachelor Thesis, „Optimierung von Modellen zur Vorhersage von Freien Hydratationsenthalpien".

[153] https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/pKa/experimental_data/NMR_microstate_determination (last accessed 22/10/2020)

[154] Frisch MJ et al (2016) Gaussian 16 Rev B.01, Gaussian Inc., Wallingford.

[155] Matsuo S, Makita T (1989) Int J Thermophys 10:885-898.

[156] https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_with_reassigned_categories/analysis_outputs/StatisticsTables/RMSE_vs_method_plot_colored_by_method_category.pdf (last accessed 2020/04/15)

[157] Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) J Comb Chem 1:55-68.

[158] https://github.com/samplchallenges/SAMPL6/blob/master/physical_properties/logP/analysis_of_extra_molecules/logP_experimental_values.csv (last accessed 2020/04/15)

[159] Slater B, McCormack A, Avdeef A, Comer JEA (1994) J Pharm Sci 83:1280-1283.

[160] Powling J, Bernstein HJ (1951) J Am Chem Soc 73:4353-4356.

[161] Spencer JN, Holmboe ES, Kirshenbaum MR, Firth DW, Pinto PB (1981) Can J Chem 60:1178-1182.

[162] Tielker N, Eberlein E, Hessler G, Schmidt KF, Güssregen S, Kast SM, *under review* (PLATZHALTER)

[163] Chemicalize 2019/05, https://chemicalize.com (last accessed 20/06/22), developed by ChemAxon (https://www.chemaxon.com, last accessed 20/06/22).

[164] Shaw DG (2005) J Phys Chem Ref Data 34:657-708.

[165] Klamt A, Eckert F, Reinisch J, Wichmann K (2016) J Comput Aided Mol Des 30:959-967.

[166] Lin B, Pease JH (2013) Comb Chem High Throughput Screen 16:817-825.

[167] Hill AP, Young RJ (2010) Drug Discov Today 15:648-655.

[168] Hölzl C, Kibies P, Imoto S, Frach R, Suladze S, Winter R, Marx D, Horinek D, Kast SM (2016) J Chem Phys 144:144104.

[169] Hölzl C, Kibies P, Imoto S, Noetzel J, Knierbein M, Salmen P, Paulus M, Nase J, Held C, Sadowski G, Marx D, Kast SM, Horinek D (2019) Biophys Chem 254:106260.

[170] Daniel I, Oger P, Winter R (2006) Chem Soc Rev 35:858-875.

[171] Meersman F, Daniel I, Bartlett DH, Winter R, Hazael R, McMillan PF (2013) Rev Mineral Geochem 75:607-648.

[172] Inagaki F, Hinrichs, KU Kubo Y, Bowles MW, Heuer VB, Hong WL, Hoshino T, Ijiri A, Imachi H, Ito M, Kaneko M, Lever MA, Lin YS, Methé BA, Morita S, Morono Y, Tanikawa W, Bihan M, Bowden SA, Elvert M, Glombitza C, Gross D, Harrington GJ, Hori T, Li K, Limmer D, Liu CH, Murayama M, Ohkouchi N, Ono S, Park YS, Phillips SC, Prieto-Mollar X, Purkey M, Riedinger N, Sanada Y, Sauvage J, Snyder G, Susilawati R, Takano Y, Tasumi E, Terada T, Tomaru H, Trembath-Reichert E, Wang DT, Yamada Y (2015) Science 349:420-424.

[173] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) J Comput Chem 25:1157-1174.

[174] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) J Comp Chem 26:1781-1802.

[175] Miyamoto S, Kollman PA (1992) J Comput Chem 13:952-962.

[176] Darden T, York D, Pedersen L (1993) J Chem Phys 98:10089-10092.

[177] Martyna GJ, Tobias DJ, Klein ML (1994) J Chem Phys 101:4177-4189.

[178] Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) J Chem Phys 103:4613-4621.

[179] Allen MP, Tildesley DJ (2017) Computer simulation of liquids. Oxford University Press, Oxford, UK.

[180] Floriano WB, Nascimento MAC (2004) Braz J Phys 34:38-41.

[181] Bandura AV, Lvov SN (2006) J Phys Chem Ref Data 35:15-30.

[182] Eberlein L (2021) PhD thesis (submitted).

Eidesstattliche Versicherung (Affidavit)

Name, Vorname
(Surname, first name)

Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum
(Place, date)

Unterschrift
(Signature)