Technische Universität Dortmund

Dissertation

Multiscale approximations of integral equation-based solvation models

Der
Fakultät für Chemie und Chemische Biologie
vorgelegt zur Erlangung des akademischen Grades
Doktor der Naturwisschenschaften
(Dr. rer. nat.)
von

Lennart Eisel
geb. am 31.01.1995
in Mülheim an der Ruhr

August 2025

1. Gutachter: Prof. Dr. Stefan M. Kast
2. Gutachter: Prof. Dr. Paul Czodrowski

Diese Arbeit wurde von Januar 2019 bis August 2025 unter der Anleitung von Prof. Dr. Stefan M. Kast in der Arbeitsgruppe Physikalische Chemie III der Fakultät für Chemie und Chemische Biologie an der Technischen Universität Dortmund angefertigt.

Contents

Danksagung

Zunächst möchte ich mich bei Herrn Prof. Dr. Stefan M. Kast für die Übernahme des Erstgutachtens und die Möglichkeit, diese Arbeit in seiner Arbeitsgruppe durchzuführen, bedanken. Herrn Prof. Dr. Paul Czodrowski danke ich für die Übernahme des Zweitgutachtens.

Außerdem möchte ich mich bei Herrn Dr. Matthias Hennemann und Herrn Prof. Dr. Timothy Clark bedanken, deren Unterstützung und Expertise dazu beigetragen haben, dass ihre Software EMPIRE erfolgreich in mein Solvatationsmodell integriert werden konnte.

Dr. Nicolas Tielker danke ich für zahlreiche Erläuterungen zu seinen Mathematica-Notebooks sowie zum SAMPL6-Papier. Lars Schumann und Stefan Maste danke ich für die gemeinsame Arbeit am Pentapeptid-Projekt. Bei Lars bedanke ich mich auch für die Durchführung der MD-Simulation des Pentapeptids und das anschließende Struktur-Sampling. Stefan danke ich ebenfalls für die Erstellung der explizit solvatisierten Pentapeptidstrukturen sowie für die produktive Diskussion der NMR-Thematik. Dr. Patrick Kibies danke ich für die vielen hilfreichen technischen und wissenschaftlichen Diskussionen sowie für das Korrekturlesen dieser Arbeit. Allen genannten sowie den weiteren Mitgliedern der Arbeitsgruppe danke ich für die angenehme Arbeitsatmosphäre.

Bei Mithil und Fynn bedanke ich mich für die Ablenkung vom Unialltag bei diversen Spieleabenden und beim Sport. Mein besonderer Dank gilt der Familie Böttcher-Pongratz, insbesondere Fini, für die Freundschaft und Unterstützung während der Pandemiezeit. Mein besonderer Dank gilt auch meinen Freunden vom Lauftreff des Hochschulsports (hier kann ich nicht alle namentlich nennen) für die gute Zeit beim Training, bei Wettkämpfen und dazwischen. Den sportlichen Ausgleich mit euch werde ich sehr vermissen.

Abschließend möchte ich mich bei meinen Eltern für die stetige Unterstützung während meines Studiums bedanken.

In so far as quantum mechanics is correct, chemical questions are problems in applied mathematics.

Henry Eyring, John Walter and George E. Kimball[1]

Despite the exciting opportunities offered by multiscale modeling, one thing we have learned during the past decade is that we should not expect quick results.

Weinan E[2]

Kurzfassung / Abstract

Die Anwendbarkeit und Kosten quantenchemischer Solvatationsmodelle, wie dem „embedded cluster reference interaction site model“ (EC-RISM) werden vor allem durch die Berechnungskosten der Elektronenstruktur, sowie des elektrostatischen Potentials bestimmt, bedingt durch die iterative Lösung dieser Verfahren und der damit einhergehenden Wiederholung der teuren Elektronenstrukturrechnung.

Hierdurch wird ihre Anwendung auf chemische Systeme geringer Größe beschränkt. In dieser Arbeit wird ein neues Solvatationsmodell, ONIOM-EC-RISM, präsentiert mit dem dieser Rechenaufwand für das bisherige EC-RISM Verfahren, durch die Einführung einer multiskalen Approximation der Elektronenstruktur, drastisch reduziert werden kann. Hierdurch erschließt sich die Möglichkeit das EC-RISM-Solvatationsmodell auch auf größere chemische Systeme, wie z.B. Proteinsysteme, anzuwenden.

Neben den für das Verständnis des ONIOM-EC-RISM-Modells relevanten theoretischen Grundlagen additiver und subtraktiver Multiskalenapproximationen, des statistisch-thermodynamischen RISM-Solvatationsmodells, sowie des methodisch verwandten ONIOM-PCM-Solvatationsmodells, werden die Theorie und technische Implementation des neuen Modells umfangreich dargestellt. Darüber hinaus wird aufgezeigt wie zuvor für das EC-RISM-Referenzmodell verwendete empirische Korrekturen in den ONIOM-EC-RISM-Kontext übertragen werden können. Hierbei wird das Größenextrapolationslimit der ONIOM-Methode ausgenutzt, wodurch Korrekturen erhalten werden, die frei von jeglichen Partitionierungsfehlern sind.

Die resultierenden Modelle sind in der Lage, die pK_a-Vorhersagequalität des EC-RISM-Referenzmodells für den pK_a-Datensatz der SAMPL6-Challenge, einem „blind prediction“ Wettbewerb zur Vorhersage thermodynamischer Größen, zu reproduzieren und teilweise zu übertreffen, während gleichzeitig die Gesamtkosten des EC-RISM-Verfahrens drastisch reduziert werden können.

Neben der Validierung des ONIOM-EC-RISM-Verfahrens anhand von pK_a-Werten wird demonstriert, wie das ONIOM-EC-RISM-Modell zusätzlich zur Vorhersage chemischer Verschiebungen eines Pentapeptidsystem verwendet werden kann. In diesem Zusammenhang wird ein neuartiger Ansatz vorgestellt der es erlaubt, pH-abhängige chemische Verschiebungen direkt aus spektroskopischen sowie pK_a-Vorhersagen zu berechnen. Dies eröffnet erstmalig die Möglichkeit der direkten Modellierung von NMR-Titrationsexperimenten auf Grundlage des EC-RISM-Solvatationsmodells.

The applicability and costs of quantum chemical solvation models, such as the "embedded cluster reference interaction site model" (EC-RISM), are mainly determined by the calculation costs of the electronic structure and the electrostatic potential, due to the iterative solution of these methods and the associated repetition of the expensive electronic structure calculation.

This limits their application to chemical systems of small size. In this work, a new solvation model, ONIOM-EC-RISM, is presented which drastically reduces the computational cost of the previous EC-RISM method by introducing a multiscale approximation of the electronic structure. This opens up the possibility of applying the EC-RISM solvation model to larger chemical systems, such as protein systems.

In addition to the relevant theoretical foundations of additive and subtractive multiscale approximations, required for the understanding of the ONIOM-EC-RISM model, the statistical-thermodynamical RISM solvation model and the methodologically related ONIOM-PCM solvation model, as well as the theory and technical implementation of the new model are presented in detail. In addition, it is shown how empirical corrections previously used for the EC-RISM reference model can be transferred to the ONIOM-EC-RISM context. Here, the size extrapolation limit of the ONIOM method is utilised, whereby corrections are obtained that are free of any partitioning errors.

The resulting models are able to reproduce and partially exceed the pK_a prediction quality of the EC-RISM reference model for the pK_a dataset of the SAMPL6 challenge, a blind prediction challenge for thermodynamic quantities, while at the same time drastically reducing the overall cost of the EC-RISM method.

In addition to the validation of the ONIOM-EC-RISM method on the basis of pK_a values, it will be demonstrated how the ONIOM-EC-RISM model can additionally be used to predict chemical shifts of a pentapeptide system. In this context, a novel approach is presented that allows pH-dependent chemical shifts to be calculated directly from spectroscopic and pK_a predictions. This opens up for the first time the possibility to directly model NMR titration experiments on the basis of the EC-RISM solvation model.

Part I.

Introduction

1. Overview

The purpose of this part is to provide a quick glance at the topics that will be discussed in this thesis and give an overview of the general structure and outline of this document. After the introduction the reader is made familiar with the theory of multiscale modelling of solutes and the solvent environment. In particular the theoretical development and technical implementation of the novel ONIOM-EC-RISM solvation model will be presented in detail. This solvation model combines the statistical-mechanical solvent description of the "embedded cluster reference interaction site model" (EC-RISM),[3] with an extrapolative multiscale approximation of the solutes electronic structure called "our own n-layered integrated molecular orbital and molecular mechanics" (ONIOM).[4,5] ONIOM-EC-RISM can be considered the core piece of this work and apart from its derivation from the "three dimensional reference interaction site model" (3D-RISM),[6–8] the second part also aims to provide a brief theoretical overview over related methods, by highlighting similarities to the ONIOM-EC-RISM methodology. This includes an overview over additive multiscale models, such as combined quantum mechanical and molecular mechanical models (QM/MM) and the "polarizable continuum model" (PCM).[9,10] Especially, it will be shown how ONIOM-PCM,[11,12] a multiscale approximation to PCM, greatly influenced the development of the ONIOM-EC-RISM model and how approximations from that model can be transferred into the RISM-context. These related methods will be discussed to an extent that aids the understanding of the considerations that need to be made for the development and implementation of the multiscale EC-RISM method.

After its extensive derivation, the ONIOM-EC-RISM solvation model will be validated and employed to investigate thermodynamic and spectroscopic properties of molecular systems. This effort will be split into two parts: The first deals with the prediction of the thermodynamic properties, in particular the prediction of macroscopic acidity constants for various systems, ranging from small molecules to larger biomolecular systems. The development of the ONIOM-EC-RISM model was in part motivated by the cost of the standard EC-RISM approach, thus the considerations that need to made to reduce these costs will be highlighted. The second part is concerned with demonstrating that this multiscale model can also be employed to predict spectroscopic parameters, in particular chemical shifts from nuclear magnetic resonance spectroscopy (NMR). In order to conclude this work, the remaining parts will summarise the findings, give an outlook and starting point for derivative works, as well as presenting additional details in the appendix.

1.1. Multiscale models in chemistry

Scientists who wish to model chemical processes are facing a dilemma. In principle, the laws that govern the microscopic behaviour of molecules are well known through the principles of quantum mechanics (QM). However, the complexity of the underlying equations and the effort required to solve them limits the size of the systems that can be studied from first principles. Although increasing computing power makes it possible to study ever larger systems using quantum mechanical methods, it is clear that these methods are not suitable for describing large-scale engineering problems or, at present, even for studying large biomolecular systems.

One way around this limitation has been the development of macroscopic models through the use of appropriate approximations. Although such an approach can reduce the overall cost of the computational process, it is often accompanied by a loss of accurate description of the microscopic dynamics of the system. An example of such model reduction from QM is molecular mechanics (MM). The classical mechanical description of the system by mapping the charge distribution onto atom-centred partial charges allows a time-resolved simulation of its dynamics up to a certain system size, but the accurate description of the electronic structure and thus its polarisation is lost. The choice of an appropriate model and its scale is therefore subject to a constant trade-off between the computability of the system and the loss of information associated with its approximation.

But what if the given chemical problem requires a combination of the accuracy of quantum mechanical methods with the more efficient description of large-scale dynamics provided by macroscopic models? In this case multiscale models may be applicable. The local nature of many chemical processes allows the expensive microscale model to be applied only to a small region of the system, while the rest is described by a macroscale model. This process can be easily illustrated using an enzyme-substrate complex. The chemical transformation of the substrate takes place in a defined region, the active site of the enzyme. In addition to the substrate molecule, the surrounding amino acids of the protein are often involved in this process. By knowing the chemical process that takes place, the change in electronic structure can be confined to a defined region. It is assumed that the rest of the enzyme is not involved in the reaction and the application of the expensive microscopic model, in this case often a suitable quantum mechanical method, can be restricted to the reactive centre. The cheaper macroscopic model is applied to the rest of the system. Due to the overall size of the systems treated, a molecular mechanical model is often chosen for this purpose. In recognition of the power of such QM/MM multiscale models, the 2013 Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel, who laid the groundwork for these models in the 1970's.[13–16]

Although QM/MM models are the most prominent types of multiscale models, the field is quite extensive and the author would like to refer the reader to the publication

"Principles of multiscale modelling" by Weinan E for a broader overview of the topic.[2]

Another important application of multiscale models in chemistry is the description of molecules in solvent environments. The considerations made above for the exemplary case of an enzyme-substrate complex are easily transferable to solute-solvent systems, and large parts of this thesis will be concerned with the theory, implementation and validation of solvation models that combine various macroscale descriptions of the solvent environment with quantum mechanics, with the intention of reducing the overall computational cost. Before proceeding with this task in the following chapters, it is useful to contextualise the ONIOM-EC-RISM model presented here by briefly familiarising oneself with previous work that employ the EC-RISM solvation model.

1.2. Previous relevant works using EC-RISM

Although the intricate details of the EC-RISM model are presented in great detail in Chapter 4, it will be advantageous to briefly familiarise the reader with its basics at this point. EC-RISM combines a quantum mechanical description of the solute with a statistical mechanical description of the solvent environment.[3] In this solvation model, the free energy of the solute is estimated by iteratively polarising its wave function by a set of point charges, which in turn are obtained from the electrostatic potential of the solute based on the 3D-RISM[6–8] solvation model.

Since its initial description by Kloss et al. in 2008, the total computational cost of the solvation model is therefore mainly determined by the cost of the electronic and solvent structure calculations performed in each iteration. These costs have remained roughly constant. However, a major improvement was presented by Jochen Heil in his doctoral thesis.[17] A drastic reduction in the number of polarising solvent charges was achieved using a Voronoi cell-based compression algorithm. As a result, the time required for the quantum mechanical code to read the charges, and thus the overall runtime of the EC-RISM code, was greatly reduced. Further cost reduction was achieved in the same work by introducing a Particle-Mesh-Ewald approach and a parallelisation strategy into the 3D-RISM code, which accelerated the solvent structure calculation.

Recently, efforts have been made to integrate EC-RISM into the quantum chemistry code ORCA,[18,19] resulting in a scheme that, similar to PCM, would recalculate the solvent structure in each iteration of the electronic structure optimisation, which is expected to reduce the overall computational cost of the solvation model. In the context of PCM, this method is referred to as a "single iteration scheme". However, this implementation is not yet available. The details of the current double iteration scheme and its differences to the single iteration scheme are further explained in chapter 4.

Otherwise, no major computational cost improvements were presented for the EC-RISM model. Therefore, a large part of this thesis will be devoted to the reduction of these costs through the development of the multiscale solvation model ONIOM-EC-

RISM, although, it should be noted that in the aforementioned work Jochen Heil presented an additive QM/MM approximation to EC-RISM. However, this approach was not designed to incorporate empirical corrections, which have been shown to be essential for accurate predictions of acidity constants and other thermodynamic properties, and the model has found little use since.

EC-RISM was first used to calculate gauche-trans equilibria for dichlormethane and pK_a shifts. Since then, the predictive power of the solvation model for thermodynamic quantities has been repeatedly tested and optimised within the SAMPL (Statistical Assessment of the Modelling of Proteins and Ligands) blind prediction challenges. The first participation was in the SAMPL2 challenge, where participants were asked to predict the free energies of solvation for a set of small molecules.[20,21]

For the next SAMPL5 challenge,[22] the organisers asked participants to calculate partition coefficients for small drug-like molecules between cyclohexane and water. Here, Tielker et al. first presented an empirical correction for the solute free energy that greatly improved the prediction results for the solvation free energies.[23] This linear correction used four correction parameters, one each for the excess chemical potential, the partial molar volume and the charge of the solute, the first two of which were obtained from the 3D-RISM solvation model. In addition, an additive offset parameter was applied.

In the EC-RISM submission[24] to the subsequent SAMPL6 pK_a prediction challenge,[25] it was demonstrated that the number of these correction parameters could be reduced to two, one for the solute charge and one for the partial molar volume. This approach was used in conjunction with a second empirical correction for the free energy difference between two adjacent protonation states of a molecule to predict acidity constants for the challenge dataset. These two corrections will also play an important role in this work and will be adapted for the ONIOM-EC-RISM solvation model.

In addition to the cited works, Nicolas Tielker's doctoral thesis[26] provides additional information on their development and the associated correction strategies.

In addition to predicting thermodynamic properties, EC-RISM has been successfully used to predict spectroscopic properties, in particular NMR parameters such as chemical shifts. These capabilities of EC-RISM were first demonstrated in a proof-of-principle study by Frach and Kast in 2014.[27] Subsequent studies focused on predicting the pressure dependence of chemical shifts, the first of which was published in 2016, again by Frach, Kast and others.[28,29] Since then, a number of EC-RISM papers have dealt with this topic.[30–33]. Further information on these topics can also be found in the doctoral theses of Roland Frach[34] and Tim Pongratz.[35]

Pressure dependence is not investigated in this work, although such calculations would be possible with ONIOM-EC-RISM. However, one of the aforementioned papers provides a methodological basis for parts of this work: Recently, Maste et al. presented a chemical shift prediction method using molecular dynamics conformer sampling based on a classical force field.[33] In addition, the influence of explicit solvent molecules on the prediction quality was investigated. Both of these approaches will be used in the later parts of this

thesis to predict chemical shifts for a pentapeptide. In addition, this work will build on the themes of these previous works by combining thermodynamic and spectroscopic calculations to directly predict the outcome of NMR titration experiments.

Before presenting the results of this newly developed method, the following chapters will familiarise the reader with the basics of multiscale models, ONIOM-EC-RISM and related work.

Part II.

Theory of multiscale solvation models

2. Multiscale approximations of the electronic structure

As summarised in Part I, the following chapters deal with the derivation of the ONIOM-EC-RISM multiscale solvation model and briefly outline the theory of related models relevant to RISM.

The derivation of these solvation models can first be divided into two hierarchically separated approximations, as depicted in figure 2.1. The aim of all multiscale approximations of molecular systems is to confine the expensive microscopic model to only those parts of the system that are directly linked with the chemical process being studied. Therefore a reasonable partitioning scheme must first be found. A given solvated system can first be partitioned into a solute with some local solvent molecules, if any, and the rest of the solvent environment. Applying a microscopic model to the first subsystem and a second macroscopic model leads to the first multiscale approximation. Modelling the solvent and solute at different scales is advantageous as it does not require cutting covalent bonds, which would introduce additional boundary conditions to the multiscale model.[2] While the macroscopic representation of the solvent environment can achieve the first cost reduction compared to microscopic, explicit solvation models, the solute is often described completely atomistically, which may limit the solute sizes that can be modelled. Examples for these types of models include QM/MM methods with explicit solvent molecules, RISM-based methods and continuum models such as PCM. The latter will be discussed in chapters 3 and 4.

In a second step, a multiscale approximation can be introduced for the solute to further constrain the costly microscale model to smaller regions of the solvated system, i.e. the active region that participates in the studied process. For QM/MM this often means that parts of the solute are added to the modelling scale of the solvent and are no longer descriped with the microscopic model. For implicit solvation models such as PCM or EC-RISM, however, a third modelling scale is added. These methods, especially in the form of QM/MM, are frequently employed to simulate extensive molecular systems, including proteins in solution.[36–38] However, they can also be utilised to examine the characteristics of small molecules, as will be exemplified later. In the following, the most common models for describing the solute, i.e. additive and subtractive multiscale models, will be introduced before their integration into implicit solvation models can be discussed in detail.

Both additive and subtractive multiscale models can be classified as domain decom-

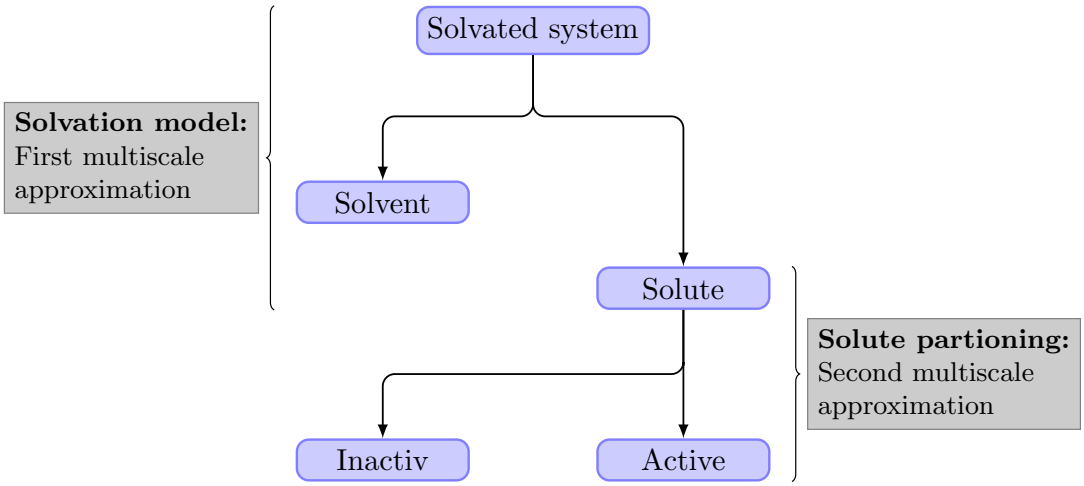

Figure 2.1.: Hierarchical multiscale modelling of a solvated system. The first step involves applying an explicit or implicit multiscale solvation model to the system, which divides it into solute and solvent. A second multiscale approximation can be applied to the solute to lower the computational costs of the model. An additive or subtractive multiscale model is utilised here to limit the use of the expensive microscale model to the chemically active region of the solute.

position methods (DDM), where the total system

$$\mathcal{A} = \bigcup_i \mathcal{A}_i \tag{2.1}$$

is divided into smaller sub-domains $\mathcal{A}_i$, which may or may not overlap.[2] The overall solution of the given problem can be approximated by the solution of the individual sub-problems. In case these nature of the problem allows it, different model scales may be applied to the sub-domains.[2]

2.1. Additive multiscale models

2.1.1. The QM/MM energy

The most prominent types of energy-based formulations of DDMs encountered in chemical modelling are two-layered, additive QM/MM models. Here the domain is divided into two subsystems

$$\mathcal{A} = \mathcal{A}_{\mathrm{QM}} \cup \mathcal{A}_{\mathrm{MM}} \tag{2.2}$$

which are defined as non-overlapping

$$\mathcal{A}_{\mathrm{QM}} = \mathcal{A} \setminus \mathcal{A}_{\mathrm{MM}}. \tag{2.3}$$

If this decomposition leads to the breaking of covalent bonds and thus to open valences, e.g. due to the QM modelling of some relevant amino acids of an enzyme binding pocket, they must be saturated by a suitable method. This means that further boundary conditions have to be defined on the QM/MM interface $\mathcal{A}_{\text{link}}$. This may include, for example, the use of link atoms. This method and other ways of modelling the interface will be discussed in section 2.1.3.

Applying a QM method as the microscopic model to the capped system $\mathcal{A}_{\text{QM}} \cup \mathcal{A}_{\text{link}}$ leads to the energy contribution E^{QM}, while the energy for the remaining system $\mathcal{A}_{\text{MM}}$ is derived using a MM-based expression

$$\begin{aligned} E_{\text{MM}} = &\sum_{\text{bonds}} k_r \left(r - r_0\right)^2 + \sum_{\text{angles}} k_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} k_\phi \left[1 + \cos\left(n\phi + \delta\right)\right] \\ &+ \sum_{\substack{\text{nonbonded} \\ \text{pairs AB}}} \varepsilon_{\text{AB}} \left[\left(\frac{\sigma_{\text{AB}}}{r_{\text{AB}}}\right)^{12} - \left(\frac{\sigma_{\text{AB}}}{r_{\text{AB}}}\right)^{6}\right] + \frac{1}{4\pi\varepsilon_0} \frac{q_{\text{A}} q_{\text{B}}}{r_{\text{AB}}}, \end{aligned} \tag{2.4}$$

which is here displayed in a generic force field form.[37] Once the solutions of the two subsystems have been found, their interaction must be considered. Methods to calculate the associated quantity E^{inter} will be presented in section 2.1.2. The total energy for the QM/MM multiscale method is thus written as

$$E^{\text{QM/MM}}\left(\mathcal{A}_{\text{QM}}, \mathcal{A}_{\text{MM}}, \mathcal{A}_{\text{link}}\right) = E^{\text{QM}}\left(\mathcal{A}_{\text{QM}} \cup \mathcal{A}_{\text{link}}\right) + E^{\text{MM}}\left(\mathcal{A}_{\text{MM}}\right) + E^{\text{inter}}\left(\mathcal{A}_{\text{QM}}, \mathcal{A}_{\text{MM}}\right). \tag{2.5}$$

Since this expression is additive, the method is called additive QM/MM.[37]

2.1.2. Coupling QM and MM subsystems

The separation of the whole system into two different subsystems and the subsequent calculation on two different scales reduces the overall computational effort and makes it possible to calculate large systems, but this also means that the interaction of the subsystems is initially neglected. The construction of the QM/MM method therefore requires finding a method that can reproduce the interaction of $\mathcal{A}_{\text{QM}}$ and $\mathcal{A}_{\text{MM}}$ from the reference description of the unpartitioned system $\mathcal{A}$ with a single QM method. The contributions to the interaction energy can be split into terms for electrostatic, van-der-Waals and covalent interactions of the subsystems[37,39]

$$E^{\text{inter}} = E^{\text{inter}}_{\text{elec}} + E^{\text{inter}}_{\text{vdW}} + E^{\text{inter}}_{\text{cov}}. \tag{2.6}$$

The focus here will be on the electrostatic part of the interaction and in particular how polarising effects of one subsystem can be mapped onto the other subsystem. The covalent term will be discussed in section 2.1.3. The van-der-Waals-interaction $E^{\text{inter}}_{\text{vdW}}$ is

often modelled by parameterised potential functions, such as a Lennard-Jones potential.[39]. These and various other approaches that have been used to approximate the contribution to $E_{\mathrm{vdW}}^{\mathrm{inter}}$ have been extensively reviewed in the literature, and the reader shall be referred to publications by Walter Thiel and other influential authors on the field of QM/MM for a detailed overview.[36,37,39–41]

Mechanical embedding

The most common methods to approximate the electrostatic interaction $E_{\mathrm{elec}}^{\mathrm{inter}}$ fall into three categories: mechanical embedding, electronic embedding and polarised embedding.[37,39] In mechanical embedding (ME), the interaction of the subsystems is modelled purely on the scale of molecular mechanics, where the charge distribution of the QM system is mapped onto atom-centred partial charges. The electrostatic part of the interaction is thus modelled at the MM scale by using a Coulomb interaction potential between the charges of both subsystems.[37] This method has the disadvantage that the mutual polarisation of the subsystems cannot be represented, since the purely MM-based description of the interaction does not allow to model the influence of the partial charges of $\mathcal{A}_{\mathrm{MM}}$ on the charge density of $\mathcal{A}_{\mathrm{QM}}$ and vice versa. Another drawback of ME is the choice of appropriate partial charges for the QM atoms and their compatibility with the MM model. The force field charges of the latter are carefully designed to give a balanced description of the potential energy surface in interaction with all other force field parameters, and are not primarily designed to accurately model the true conformation dependent charge distribution and energetics of the system. It is therefore not justifiable to derive QM charges with a model other than the one used to develop the force field.[37,41] In addition to this drawback, the point charges for QM atoms are not easily updated, for example due to electronic structure changes during chemical reactions, as this would lead to discontinuities in the potential energy surface.[37] The mechanical embedding scheme is therefore considered somewhat outdated.[39]

Electronic embedding

In contrast, electronic embedding (EE) is a more sophisticated approach to modelling the interaction of the two subsystems by considering the polarising effect of $\mathcal{A}_{\mathrm{MM}}$ on the subsystem $\mathcal{A}_{\mathrm{QM}}$. This is done by embedding the cluster of QM atoms in the charge field of the MM atoms. The electronic Hamiltonian of the QM calculation in atomic units

$$\hat{H}_{\mathrm{e}}^{\mathrm{EE}} = \hat{H}_{\mathrm{e}} - \sum_{i}^{N_{\mathrm{e}}} \sum_{j \in \mathcal{A}_{\mathrm{MM}}} \frac{q_j}{\|\mathbf{r}_i - \mathbf{R}_j\|} + \sum_{k}^{N_{\mathrm{n}}} \sum_{j \in \mathcal{A}_{\mathrm{MM}}} \frac{q_j Z_k}{\|\mathbf{R}_k - \mathbf{R}_j\|} \tag{2.7}$$

therefore includes an additional electrostatic term modelling the potential field of the MM point charges q_j. Here $\hat{H}_{\mathrm{e}}$ is the electronic Hamiltonian of the respective QM

method without those additional charges and the number of electrons and nuclei in $\mathcal{A}_{\mathrm{QM}}$ is represented by N_{e} and N_{n}.[39] This has the advantage that for the QM atoms there is no need to derive a point charge model that fits the force field parameters. In addition, the electronic part of the interaction energy is now calculated at the QM level rather than at the molecular mechanical level as in ME, which provides a more advanced description of the electrostatic effects in the system and is generally considered to give reasonable results.[41] Furthermore, the inclusion of the partial charges in the Hamiltonian is efficient and the charges are readily available from force fields.[41] Most modern quantum chemistry codes, such as Gaussian[42] or ORCA,[18,19] include QM/MM models that already allow electronic embedding or the specification of additional point charges, so that such a method can be easily constructed. However, apart from all these improvements of EE over ME, there is still the problem that the MM partial charges are parameterised in interaction with all other force field parameters and do not necessarily reflect the true charge distribution of $\mathcal{A}_{\mathrm{MM}}$ one would obtain from a first principles model.[37] Furthermore, EE allows the description of polarisation effects on the QM system, but not mutual polarisation, which would be present in the QM description of the whole system. These shortcomings arise from the use of classical force fields, which do not allow for polarisation effects at all due to the use of rigid point charges.

Polarised embedding

Polarised embedding (PE) schemes try to solve these problems by switching to a description of $\mathcal{A}_{\mathrm{MM}}$ that also allows for a mutual polarisation by the charge distribution of $\mathcal{A}_{\mathrm{QM}}$.[43] In fact the first QM/MM model by Warshel and Levitt makes use of PE.[16,41,43] Commonly, polarisable force fields based on Drude oscillators[43–47] or induced point dipoles[16,43,48–52] are used for this purpose, but other approaches such as force fields derived from first-principles have also gained attention in the last decade.[43,53,54]

Although the basic problem of mutual polarisation in additive QM/MM approaches has been known since their first implementation[16] and great hopes are placed on the future development of these methods, these approaches are currently not as advanced as fixed-charge approaches.[39,43] Furthermore, the self-consistent solution of the mutual polarisation of both subsystems leads to further disadvantages compared to EE approaches, as the polarisation equations required for PE schemes must be solved in each SCF step, increasing the overall computational effort and requiring direct integration of this multiscale method into the QM code.[41,43] As Bondanza et al. point out in their review, a necessary condition for the application of QM/MM-PE approaches is that they must first be available as robust, ready-to-use implementations, and only in this way can they be made accessible to a wider community.[43] Although these approaches are also interesting for advanced problems such as exited state properties and spectroscopy, this requirement prevents their application to the multiscale RISM solvation models that are developed in this work, and we will revert to simpler but more robust EE approaches.

Another common problem in both electronic and polarised embedding is overpolarisation. The truncation of the system for the QM calculation and the resulting macroscopic description of the omitted nuclei and electrons, by point charges leads to the neglect of Pauli repulsions. The electronic density close to the QM-MM boundary can therefore be overpolarised when interacting with positively charged MM point charges, since the interaction is purely attractive. This effect, known as "electron spill out", can be particularly pronounced if the basis set is flexible enough and may require the use of additional model potentials that attempt to recover the effects of Pauli repulsion.[39,55] Another, and likely simpler, method to solve this issue is by moving the QM-MM boundary further away from the chemically active centre. This can be achieved by including larger parts of the complete system in the QM calculation, thereby reducing the electrostatic impact of the overpolarised boundary on the inner system. This, of course, results in an increase in computational cost for the quantum mechanical calculation.[39]

2.1.3. Modelling the QM-MM interface

The final term, $E_{\text{cov}}^{\text{inter}}$, which contributes to the subsystems' interaction energy E_{inter} from equation 2.6, originates from covalent bonds that cross the boundary between the QM and MM zones, i.e. one bond partner is in $\mathcal{A}_{\text{QM}}$, and the other is in $\mathcal{A}_{\text{MM}}$. If the system can be partitioned in such a way that no covalent bonds cross the interface, the energy contribution will be zero. This applies to straightforward QM/MM solvation models where the solute is modelled at the QM level, and the solvent environment is exclusively described at the MM level. This is rarely the case, particularly, when the solute becomes larger, modelling at the QM scale becomes impossible, and a second multiscale approach is needed, as shown in figure 2.1. Therefore, while developing QM/MM methods, finding a way to handle open valencies caused by cutting bonds that cross the QM/MM interface becomes crucial.

One common approach to handle this problem is by employing link atoms (LA), which replace the bonding partner in $\mathcal{A}_{\text{MM}}$, that is not included in the QM calculation. For consistency with later discussions about link atom schemes in the context of subtractive multiscale models, this atom will be refered to as 'link-atom-host' (LAH) and its bonding partner in $\mathcal{A}_{\text{QM}}$ as 'link-atom-connection' (LAC).[5] Hydrogen link atoms are frequently used to saturate the dangling bond, but the usage of other elements or monovalent groups to recover the electronic distribution of the original bond has also been explored.[56] This approach to the QM/MM boundary has the advantage of a straightforward implementation as the link atoms may be treated just like any other atom in the QM calculation.

Placing a single linked atom without additional constraints also has its drawbacks. It would increase the number of degrees of freedom by three, making geometry optimisation more difficult. It would also pose a challenge as to where to place the bonded atom during these optimisations. One way to solve this issue is by empirically determining

the coordinates of the link atom, based on the nuclear coordinates $\mathbf{R}_{\mathrm{LAH}}$ and $\mathbf{R}_{\mathrm{LAC}}$ of LAH and LAC.[5,57,58] The link atom position

$$\mathbf{R}_{\mathrm{LA}} = \mathbf{R}_{\mathrm{LAC}} + g\,(\mathbf{R}_{\mathrm{LAH}} - \mathbf{R}_{\mathrm{LAC}}) \tag{2.8}$$

is commonly positioned on the LAH-LAC axis by scaling the bond length with a scaling factor g. Therefore, the number of degrees of freedom in the whole system remains unchanged for the optimisation problem. The scaling factor can simply be set to 1 to position the bond atom at the LAH coordinates, or it can be chosen to give a more reasonable geometries using tabulated values for the ratio of LAC-LA and LAC-LAH bond distances,[56,57] which is the method applied throughout this work.

Aside from modelling covalent bonds across the QM-MM interface using the link atom scheme outlined previously or comparable approaches,[59,60] alternative strategies such as the pseudo-bond[61,62] or frozen orbital methods[63,64] have also been developed.[39]

So far, multiscale models have been discussed that combine a quantum mechanical method as a microscopic model and a molecular mechanical method as a macroscopic model. However, what happens if the accuracy of the MM-based method is not sufficient to describe the dynamics of the system with the desired precision, or if one wishes to combine any two methods in general? In such situations, it may be preferable to use a second QM or semiempirical-QM (SQM) method as the macroscopic model. In the early 2000s, strictly additive multiscale models were discussed, which may allow the combination of different methods.[56,65] However, they often required direct modifications of the electronic structure codes and produced limited literature since. In the meantime, subtractive methods such as the ONIOM family of models[56,66–68] have gained popularity and are now among the most widely used methods for combining any level of theory.[5] The following sections will discuss the theory behind these multiscale models, as well as their differences and potential benefits over additive models.

2.2. Subtractive multiscale models

2.2.1. IMOMM and IMOMO

We will start the discussion of subtractive multiscale methods with the "integrated molecular orbital molecular mechanics" (IMOMM) model of Maseras et al.,[66] since it tries to achieve the same thing in principle by similar means as the additive QM/MM methods described in the previous sections: reducing the overall computational effort by combining a QM method with a MM method and restricting the former to a smaller part of the total system. In the context of the IMOMM method, as well as other members of the ONIOM family of methods, this subset of atoms $\mathcal{A}_{\mathrm{QM}}$ with the associated link atoms $\mathcal{A}_{\mathrm{link}}$ is collectively referred to as the *model* system $\mathcal{A}_{\mathrm{m}}$, while the complete system is referred to as the *real* system $\mathcal{A}_{\mathrm{r}}$.

To obtain the total energy of the system with the IMOMM method, first the MM energy of the *real* system is calculated and added to the QM energy of the *model* system. In contrast to additive QM/MM methods, the two sets of atoms on which the MM and QM calculations are performed now overlap, since the *real* set is a superset of the *model* system. To correct for double counting of energy contributions from atoms present in both the *model* system and the *real* system, the MM energy is subtracted, giving the expression

$$E^{\mathrm{IMOMM}}\left(\mathcal{A}_{\mathrm{r}}, \mathcal{A}_{\mathrm{m}}\right) = E^{\mathrm{MM,real}}\left(\mathcal{A}_{\mathrm{r}}\right) + E^{\mathrm{QM,model}}\left(\mathcal{A}_{\mathrm{m}}\right) - E^{\mathrm{MM,model}}\left(\mathcal{A}_{\mathrm{m}}\right). \tag{2.9}$$

Unlike equation 2.5, the IMOMM expression does not feature an explicit interaction term, as the $E^{\mathrm{MM,real}}$ term already includes the interaction of the QM atoms with the remaining systems. Therefore, the interaction is solely modelled at the MM level.

The next stage in the evolution of subtractive multiscale methods was the creation of the "integrated molecular orbital + molecular orbital" (IMOMO) method, capable of combining a quantum mechanical approach with a less costly QM or SQM calculation.[67] Similarly to equation 2.9, the complete energy of the IMOMO method is represented by

$$E^{\mathrm{IMOMO}}\left(\mathcal{A}_{\mathrm{r}}, \mathcal{A}_{\mathrm{m}}\right) = E^{\mathrm{QM2,real}}\left(\mathcal{A}_{\mathrm{r}}\right) + E^{\mathrm{QM,model}}\left(\mathcal{A}_{\mathrm{m}}\right) - E^{\mathrm{QM2,model}}\left(\mathcal{A}_{\mathrm{m}}\right). \tag{2.10}$$

Here the second QM or SQM method is abbreviated as QM2. The form of this equation may indicate why subtractive models combining two QM methods are more commonly applied than additive ones. The absence of an explicit interaction term makes it possible to combine two quantum mechanical methods from different software packages into one multiscale method without having to make extensive modifications to their mathematical definition or technical implementation.[5] While there may be additional contributing factors, it can also be argued that the implementation of subtractive models in the Gaussian software[42,56] has made these models more accessible, potentially playing an additional role in their current popularity.

2.2.2. ONIOM

The methods presented so far divide the system into two layers and approximate the total energy by two model scales. In 1996, Morokuma and coworkers generalised these subtractive methods into "our own n-layered integrated molecular orbital and molecular mechanics" (ONIOM) model, which allows any number of methods and shells to be combined.[68] Since its first publication, ONIOM has become a popular subtractive model and will be used in later parts of this work to develop an EC-RISM-based multiscale method.

For two- and three-layer approximations, ONIOM introduces generic names for the method and layer combinations in order to generalise the method. Throughout this work, the notation from Chung et al's review of ONIOM theory[5] is used: In a two-layer

approximation, the accurate but more expensive method is called *high*, while the second cheaper method is called *low*. As within the context of IMOMM and IMOMO, ONIOM makes use of the full set of atoms, the "*real* system", and a subset of atoms and link atoms that are investigated with the *high* method. This subset is referred to as the "*model* system". The three sub-calculations are thus labelled "*low,real*", "*high,model*" and "*low,model*" or *lr*, *hm*, and *lm* for brevity in the following equations and throughout this work. To maintain consistency of notation and avoid confusion with the *hm* calculation, which shares the same level of theory, the *high* reference calculation on the *real* system is referred to as "*high,real*" or *hr*. Often a two-layer ONIOM approximation is written as ONIOM2(*high*:*low*).

The system's total energy in the ONIOM approximation is expressed, in analogy to equations 2.9 and 2.10, as

$$E^{\mathrm{ONIOM2}}\left(\mathcal{A}_{\mathrm{r}}, \mathcal{A}_{\mathrm{m}}\right) = E^{\mathrm{lr}}\left(\mathcal{A}_{\mathrm{r}}\right) + E^{\mathrm{hm}}\left(\mathcal{A}_{\mathrm{m}}\right) - E^{\mathrm{lm}}\left(\mathcal{A}_{\mathrm{m}}\right). \tag{2.11}$$

As the set of atoms on which individual sub-calculations are performed is sufficiently defined through the two-letter abbreviations, the sets in the function arguments will be omitted from now on and only explicitly shown if necessary.

The deviation from the *hr*-reference defines the error E^{Err} in an ONIOM2 calculation

$$E^{\mathrm{hr}} = E^{\mathrm{ONIOM2}} + E^{\mathrm{Err}}. \tag{2.12}$$

It is worth mentioning that both the additive QM/MM and ONIOM method are commonly employed for predicting relative energies between two states as this can potentially cancel out or at least diminish the approximation error. Approximations to the total energy of the system may be achieved using other domain decomposition techniques such as fragmentation methods.[5,53,69]

Embedding schemes

In principle, embedding schemes for ONIOM calculations are similar to those formulated for additive QM/MM methods in section 2.1.2 and shall only be outlined briefly. The most basic approach is mechanical embedding, which requires no additional modifications to the ONIOM method presented so far. Although derived from multiscale models where molecular mechanics is used for the macroscopic model scale, the term mechanical embedding is used for ONIOM methods regardless of the choice of the *low* level of theory. Within ME the energies of the sub-calculations are calculated independently and added according to equation 2.11. In the resulting extrapolated energy, the interactions between the *model* system and the remaining environment are modelled at the *low* level of theory by means of the *lr* calculation, as it describes all atoms within the system and their interaction. In contrast, additive models model the interaction of the two different subsets of atoms at the level chosen for the interaction term.

Therefore, when an ONIOM-ME approach is used in conjunction with a *low* level of theory that allows for mutual polarisation of atoms, e.g. a second QM method, the polarisation between the *model* system and the remaining chemical environment is implicitly accounted for within the *lr* calculation. However, as with additive models, the polarisation is not modelled at the *high* level of theory, and the atoms within the *model* system sub-calculations may not be polarised as in the *hr* calculation, as the interaction with the remaining system is omitted for *hm* and *lm* when using ME.

Electrostatic embedding can be achieved by including the charge distribution of the surrounding system, mapped onto point charges, in both *model* system calculations. The inclusion of partial charges allows the interaction of the *model* system with the remaining atoms to be described at the *high* level of theory, which is not possible with ONIOM-ME. In the case of an ONIOM(QM:MM) calculation, the force field charges can be used for this purpose, with all the drawbacks already described in section 2.1.2. For calculations where a second QM method or a semi-empirical method replaces the MM level of theory, the *low*-level charge distribution can also be mapped onto point charges using suitable partial atomic charges.

As electrostatic interactions, described by adding point charges to the calculations of the *model* system, are also included in the *lr* calculation, they must be excluded from the overall ONIOM energy to prevent double counting. This is accomplished by subtracting the two *model* system energies (*hm* and *lm*) in the expression for the ONIOM extrapolation, requiring no further ad-hoc corrections. The electrostatic term added to the Hamiltonians of the *model* system calculations is identical to that presented for the additive QM/MM method, although Vreven et al. scale the point charges near the bond atoms in their first derivation of the ONIOM(QM:MM)-EE method.[58] This is done to avoid over- or undercounting of interactions already included in the bond terms of the force field, as well as overpolarisation of the environment of the bond atoms.

As with additive QM/MM models, efforts have been made to extend electrostatic embedding schemes to polarisable embedding,[5] although to a lesser extent. The development of PE in additive models was motivated by the shortcomings of EE, which does not allow mutual polarisation of the QM and MM subsystems, as described in section 2.1.2. This problem can be transferred to the context of ONIOM(QM:MM) calculations, where the use of partial charges from the force field in the *hm* calculation allows polarisation of the wave function, but the charge distribution of the *model* system does not allow polarisation of the environment due to the static force field based description of its charge distribution. However, in contrast to additive QM/MM, ONIOM allows the simple substitution of the MM level of theory with a second QM, SQM or polarisable force field method capable of modelling the polarisation of the system. When such a method is used, the polarisation of the *real* system is already described by the *lr* calculation at this *low* level of theory, and possibly by the interaction of the *model* system with the EE-charges and there may be no need to go beyond a simple ME or EE scheme.

In summary, in ONIOM-ME the interaction between the subset of atoms contained in the *model* system and the remaining atoms not described by the *model* system is modelled at the *low* level of theory. In ONIOM-EE, the polarisation effect of the environment on the *model* system is also modelled at the *high* level of theory by including point charges in the *model* system calculations derived at the *low* level of theory. PE schemes attempt to model the influence of the *hm* charge density on the rest of the system.

Such an ONIOM-PE scheme has been developed by Caprasecca et al. where a QM method and a polarisable force field are combined.[70] The authors state that in their PE scheme the *model* system is polarised by MM charges as well as MM induced dipoles and that these dipoles are able to respond to the *hm* electron density. The additional energy contributions due to the PE scheme, i.e. electrostatic and polarisation terms, are evaluated within the *hm* calculation. To avoid double counting, these energy contributions are switched off for the force field contributions, i.e. in the *lr* and *lm* calculations. This model and a later improvement[71] have been used to study dyes inserted into DNA. Otherwise, the literature on ONIOM-PE models is rather limited.

Diagrammatic representation and multiple layers

ONIOM and its related subtractive methods are often referred to as extrapolative methods because they attempt to approximate the *hr* reference by two different extrapolations, firstly an extrapolation from the *low* to the *high* level of theory, and secondly a size extrapolation from the *model* to the *real* system.[5] This process is illustrated in figure 2.2.

All possible ONIOM sub-calculations can be plotted on a grid or matrix with the level of theory on one dimension and the system size or respective subset of $\mathcal{A}_r$ as the other dimension, i.e. for a two-layer ONIOM calculation, *low* and *high* on one axis and *model* and *real* on the other. These are ordered with increasing level of theory and system size. There are therefore three grid points representing the ONIOM sub-calculations and one for the *hr* reference, which is the extrapolation target. From this diagrammatic representation one can construct the mathematical representation of the ONIOM extrapolation from eq. 2.11 by adding the sub-calculations of the diagonal (*hm* and *lr*) and subtracting the *lm* calculation from the subdiagonal. In figure 2.2, this is shown by following the path from *hm* to *lm* and *lr* and using the given addition or subtraction operators to connect the respective quantities.

This representation is particularly useful when generalising the ONIOM method to more than two layers. In a three-layer approach, noted as ONIOM3(*high*:*medium*:*low*), a third, *intermediate* system is added, which is treated in an analogous way to the *model* system, i.e. by defining a subset of atoms $\mathcal{A}_i$ from the *real* system and saturating open valences with link atoms or another suitable boundary scheme. The *model* system is

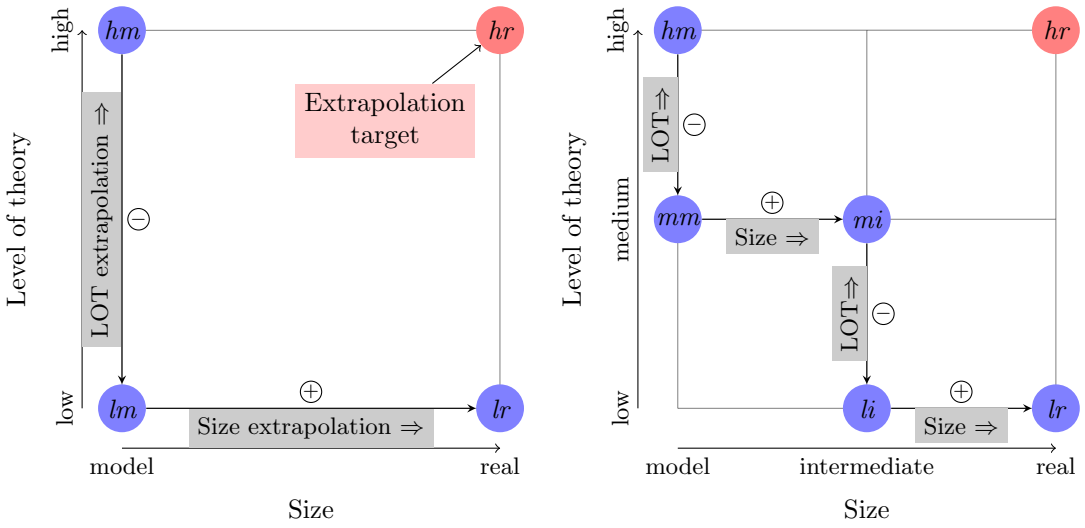

Figure 2.2.: Illustration of the ONIOM extrapolation for a two-layer system (left) and a three-layer system (right). The ONIOM method combines an extrapolation from one level of theory to a more costly level and a size extrapolation from the *model* system and *intermediate* system to the *real* system (highlighted in grey). Calculations required for each ONIOM extrapolation are highlighted as blue nodes. The associated mathematical expression is obtained by following the path connecting these nodes with their respective signs.

now defined as a subset of the *intermediate* system such that

$$\mathcal{A}_m \subseteq \mathcal{A}_i \subseteq \mathcal{A}_r. \tag{2.13}$$

In addition, a third level of theory, called *medium*, is added. The cost and accuracy of this method are generally between those of *low* and *high*. The total energy expression of the ONIOM3 method can be easily constructed diagrammatically as shown in figure 2.2. From the total of nine grid points, only the sub-calculations from the diagonal and subdiagonal are required for the extrapolation. As with the ONIOM2 approach, the energies of the subdiagonal are subtracted from the summed up energies of the diagonal, giving

$$E^{\mathrm{ONIOM3}} = E^{\mathrm{lr}} + E^{\mathrm{mi}} - E^{\mathrm{li}} + E^{\mathrm{hm}} - E^{\mathrm{mm}}. \tag{2.14}$$

As before, the newly introduced sub-calculations are abbreviated as *li*, *mi* and *mm*.

The ONIOM3 approximation can also be understood as a series of ONIOM2 extrapolations by considering the points in the upper part of the grid. Firstly, *hm*, *mm* and *mi* can be used to extrapolate to the "*high,intermediate*" (*hi*) grid point

$$E^{\mathrm{hi}} = E^{\mathrm{mi}} + E^{\mathrm{hm}} - E^{\mathrm{mm}}, \tag{2.15}$$

secondly, the grid point "*medium,real*" (*mr*) can be obtained by means of *mi*, *li* and *lr*

$$E^{\mathrm{mr}} = E^{\mathrm{lr}} + E^{\mathrm{mi}} - E^{\mathrm{li}}. \tag{2.16}$$

and the final two level extrapolation from *hi*, *mi* and *mr* gives

$$E^{\mathrm{ONIOM3}} = E^{\mathrm{mr}} + E^{\mathrm{hi}} - E^{\mathrm{mi}}, \tag{2.17}$$

which is identical to equation 2.14 after substituting the energy expression for *hi* and *mr* from the previous equations.

Three layer ONIOM extrapolations were first used by Sevensson et al. to study Diels-Alder reactions by employing various QM and MM method combinations.[68] ONIOM3 is commonly used when the application of the *medium* level of theory to the *real* system is too expensive or not required, as a less expensive method is able to capture the effects of the environment on the *intermediate* system.

The ONIOM extrapolation is not limited to two or three layers, but can be generalised to any number of layers. The expression for the total energy of such an ONIOMn calculation combing n layers and n levels of theory is given by

$$E^{\mathrm{ONIOM}n} = \sum_{i=1}^{n} E^{i,n+1-i} - \sum_{i=2}^{n} E^{i,n+2-i} \tag{2.18}$$

and requires a total of $2n-1$ sub-calculations.[5,68] Again, the combination of level of theory and system is symbolised by the superscript of each energy and the first sum represents the grid points on the diagonal of the $n \times n$ grid, while the subdiagonal grid points lead to the energy contributions in the second sum. Apart from their theoretical description within the ONIOMn generalisation, $n > 3$ extrapolations are rarely used, with two- and three-shell approximations constituting the majority of ONIOM methods described in the literature and implemented into quantum chemistry codes.[5]

Generalisation of the ONIOM notation

The ONIOM extrapolations mentioned above are all concerned with extrapolating to the total energy of the system. In the later derivation of the ONIOM-EC-RISM method it will be useful to apply the ONIOM extrapolation also to other quantities such as the electrostatic potential or in general to any quantity x. Therefore in the following a generalised notation of the ONIOM method will be presented, which also allows to shorten the lengthy expressions often encountered when dealing with this theory.

The general idea behind this notation is that the ONIOM extrapolation is a function that maps the results from the sub-calculations to the extrapolated result, which can also be rewritten as a vector dot product. For a two-layer ONIOM calculation of a quantity x,

we can collect the results of all sub-calculations into a vector $\mathbf{x}^{(2)} = \left(x^{\mathrm{lr}}, x^{\mathrm{hm}}, x^{\mathrm{lm}}\right)^{\intercal}$ and define the extrapolation with a coefficient vector $\mathbf{c}^{(2)} = (1, 1, -1)^{\intercal}$ as the dot product

$$\Omega^{(2)}(\mathbf{x}^{(2)}) \coloneqq \mathbf{c}^{(2)}\mathbf{x}^{(2)} = x^{\mathrm{lr}} + x^{\mathrm{hm}} - x^{\mathrm{lm}}, \tag{2.19}$$

which will be referred to as the ONIOM extrapolation. The superscript "(2)" denotes the order $n = 2$ of the ONIOM2 extrapolation. As the focus of this work is primarily on ONIOM2 extrapolations, this superscript will be omitted for brevity unless necessary. Therefore, Ω now refers to a two-layer ONIOM extrapolation. The expression for the ONIOM2 energy from eq. 2.11 can thus be shortened to

$$E^{\mathrm{ONIOM2}} = \Omega\left(\mathbf{E}\right). \tag{2.20}$$

There are a total of six permutations for the vector $\mathbf{x}$, which give five additional formulations of the ONIOM extrapolation

$$\Omega_{\mathrm{I}}(\mathbf{x}_{\mathrm{I}}) \coloneqq \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \begin{pmatrix} x^{\mathrm{hm}} \\ x^{\mathrm{lr}} \\ x^{\mathrm{lm}} \end{pmatrix}, \Omega_{\mathrm{II}}(\mathbf{x}_{\mathrm{II}}) \coloneqq \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} x^{\mathrm{lr}} \\ x^{\mathrm{lm}} \\ x^{\mathrm{hm}} \end{pmatrix}, \ldots, \Omega_{\mathrm{V}}(\mathbf{x}_{\mathrm{V}}) \tag{2.21}$$

that all yield identical results to Ω. For the purpose of this work, the order of $\mathbf{x}$ given in eq. 2.19, i.e. *lr*, *hm*, *lm*, is used because it reflects the order in which the sub-calculations are performed in the ONIOM-EC-RISM method as well as in quantum chemistry codes such as Gaussian16.[42]

The notation can be easily adapted for three-layer calculations by defining

$$\Omega^{(3)}(\mathbf{x}^{(3)}) \coloneqq \mathbf{c}^{(3)}\mathbf{x}^{(3)} = x^{\mathrm{lr}} + x^{\mathrm{mi}} - x^{\mathrm{li}} + x^{\mathrm{hm}} - x^{\mathrm{mm}} \tag{2.22}$$

with $\mathbf{x}^{(3)} = \left(x^{\mathrm{lr}}, x^{\mathrm{mi}}, x^{\mathrm{li}}, x^{\mathrm{hm}}, x^{\mathrm{mm}}\right)^{\intercal}$ and $\mathbf{c}^{(3)} = (1, 1, -1, 1, -1)^{\intercal}$ or any other permutation of $\mathbf{x}^{(3)}$ and corresponding order of $\mathbf{c}^{(3)}$ as described above.

For the sake of completeness, the generalisation to n-layers will also be outlined in the following. As in the diagrammatic representation in figure 2.2, the set of all possible sub-calculations can be arranged in an $(n \times n)$-matrix

$$\mathbf{X}^{(n)} = \begin{pmatrix} x_{11} & & & \cdots & x_{1n} \\ x_{21} & x_{22} & & & \\ & x_{32} & x_{33} & & \\ \vdots & & \ddots & \ddots & \\ x_{n1} & & & x_{nn-1} & x_{nn} \end{pmatrix}, \tag{2.23}$$

where a single element x_{ij} represents a calculation with method i on system j. From the subset of the $2n - 1$ sub-calculations actually needed for the ONIOMn extrapolation, there are n diagonal elements and $n - 1$ subdiagonal elements. From the vector of

the diagonal elements $\mathbf{x}_{+1} = (x_{11}, \ldots, x_{nn})^\intercal$ and the vector of the subdiagonal elements $\mathbf{x}_{-1} = (x_{21}, \ldots, x_{nn-1})^\intercal$, the vector $\mathbf{x}^{(n)} = (\mathbf{x}_{+1}, \mathbf{x}_{-1})^\intercal$ can be constructed by concatenation. Similarly, the coefficient vector $\mathbf{c}^{(n)} = (\mathbf{1}_n, -\mathbf{1}_{n-1})^\intercal$ is constructed by appending two vectors that are equal in length of $\mathbf{x}_{+1}$ and $\mathbf{x}_{-1}$. Here $\mathbf{1}_n$ represents a column vector of length n filled with ones. This finally gives the ONIOM extrapolation for n layers

$$\Omega^{(n)}(\mathbf{x}^{(n)}) = \mathbf{c}^{(n)}\mathbf{x}^{(n)} = \begin{pmatrix} \mathbf{1}_n \\ -\mathbf{1}_{n-1} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{+1} \\ \mathbf{x}_{-1} \end{pmatrix} = \sum_{i=1}^{n} x_{i,i} - \sum_{j=1}^{n-1} x_{j+1,j}. \tag{2.24}$$

Again, the permutation of $\mathbf{x}_{+1}$ or $\mathbf{x}_{-1}$, or more generally the order of $\mathbf{x}^{(n)}$, is irrelevant as long as a diagonal element is assigned a factor of $+1$ and a subdiagonal element is assigned a factor of -1 from $\mathbf{c}^{(n)}$.

It is easy to see that $\Omega^{(n)}$ is a linear map, since it satisfies additivity

$$\Omega^{(n)}(\mathbf{x}^{(n)} + \mathbf{y}^{(n)}) = \Omega^{(n)}(\mathbf{x}^{(n)}) + \Omega^{(n)}(\mathbf{y}^{(n)}) \tag{2.25}$$

with a second vector $\mathbf{y}^{(n)}$ of the same composition as $\mathbf{x}^{(n)}$, as well as homogeneity

$$\Omega^{(n)}(\lambda\mathbf{x}^{(n)}) = \lambda\Omega^{(n)}(\mathbf{x}^{(n)}) \tag{2.26}$$

with a scalar λ. This simplifies and shortens some of the derivations of quantities within ONIOM theory that will be encountered in the following chapters.

A final note on the nomenclature of additive and subtractive models: As can be seen from the theory outlined above, the ONIOM family of methods is capable of combining QM and MM models as well as other combinations of methods. Therefore, it is common in the literature to refer to the ONIOM method as QM/MM or, in the case of two QM (or SQM) methods being combined, as QM/QM. When both additive and subtractive models are discussed, this can of course lead to confusion between the two. Therefore, in this work, additive models will be referred to exclusively as QM/MM or explicitly as additive QM/MM, and subtractive models will be referred to by their respective method names, i.e. ONIOM in most cases.

3. PCM and its ONIOM approximation

From both QM/MM and ONIOM methods described in the last chapter it is already possible to construct simple multiscale solvation models where the solute or its chemically active part is modelled on the microscopic scale and the rest of the system and the surrounding solute is modelled on the macroscopic scale. These types of explicit solvation models are common and are used, for example, to study protein systems, but they still require the solvent to be modelled explicitly, which adds to the overall cost of the calculation. In the following chapters, two solvation models are presented that allow the computational cost of the solvent environment to be reduced. In this chapter, the "polarisable continuum model" (PCM)[9,10] and its multiscale approximation ONIOM-PCM[11,12] are introduced, which will lay the foundation for understanding the approximations introduced for the ONIOM-EC-RISM solvation model, which will be discussed in the last chapter of this part.

3.1. General outline of the PCM approach

In PCM, the solvated system is divided into the solute inside a cavity and the surrounding solvent. The former is modelled as a void with a relative permittivity $\varepsilon_\mathrm{i} = 1$ and the latter as a continuous, homogeneous medium with a solvent-specific ε_e.[10] The solute inside the cavity can be modelled using single or multiscale methods, as will be shown later for the ONIOM-PCM model.[11,12] Polarisation of the solute due to interaction with the solvent environment is achieved by placing point charges on the cavity surface, called apparent surface charges (ASC). The size of the ASC is in turn determined by the solute charge density, allowing mutual polarisation of the solvent and solute.

The term PCM does not refer to a single code or method, but rather to a family of methods that have arguably gained popularity through their implementation in the quantum chemistry software Gaussian. The first PCM method, now called the "dielectric continuum polarisable model" (DPCM), was first published in 1981 by Miertuš et al.[9] and has since been reformulated several times,[10] most notably as the "integral equation formalism polarisable continuum model" (IEFPCM).[72] The theory of the latter is outlined below, as it is also the PCM variant used by Vreven et al. for the ONIOM-PCM method. The structure and notation of the following sections generally follow those of the review on PCM by Tomasi et al.,[10] which is recommended to the reader for an in-depth review of the topic.

3.2. IEFPCM

In PCM, the free energy of the system at infinite dilution

$$G = G_{\mathrm{el}} + G_{\mathrm{rep}} + G_{\mathrm{dis}} + G_{\mathrm{mM}} + G_{\mathrm{cav}} \tag{3.1}$$

is phenomenologically decomposed into electrostatic, repulsion, dispersion and molecular motion contributions, as well as the energy required to form the solvent cavity.[72] This free energy is referenced to a state consisting of the unperturbed pure solvent at given temperature and the non-interacting, i.e. infinitely separated, electrons and nuclei of the solute.[72] The summary presented here focuses on the electrostatic contribution G_{el} within the IEFPCM model. For more advanced information on the remaining terms in the energy decomposition and the PCM class of methods, the reader is referred to ref. [10]. After the preliminaries, this chapter first discusses the calculation of the ASC from the solute charge density and then outlines the self-consistent integration of these charges into the Hartree-Fock formalism before moving on to the ONIOM-PCM solvation model.

3.2.1. Calculating the apparent surface charges

Since there is mutual polarisation between the solute charge distribution ρ_{M} and the dielectric continuum, the solute-solvent interaction potential must be determined in an iterative process. This interaction potential is often referred to as the solvent reaction field, although it is argued that the term solvent reaction potential is more accurate as it is the additional potential that needs to be taken into account in the solute Hamiltonian.[10] Tomasi et al. begin their derivation of the IEFPCM by first finding a solution to the classical electrostatic problem and thus the additional potential derived from the ASC and then inserting it into the HF formalism.[10]

The basic electrostatic problem

The derivation starts with the general Poisson equation

$$-\vec{\nabla}\left[\varepsilon(\mathbf{r})\vec{\nabla}V(\mathbf{r})\right] = 4\pi\rho_{\mathrm{M}}(\mathbf{r}) \tag{3.2}$$

with the electrostatic potential V. This equation can be simplified to

$$-\nabla^2 V(\mathbf{r}) = 4\pi\rho_{\mathrm{M}}(\mathbf{r}) \tag{3.3}$$

for the volume C inside the cavity and

$$-\varepsilon\nabla^2 V(\mathbf{r}) = 0 \tag{3.4}$$

for the outside of C. Here ε is used as a constant, and more specifically $\varepsilon = 1$ inside the cavity.[73] It is assumed that ρ_{M} is zero outside the cavity.[10]

The electrostatic potential is decomposed into two terms

$$V(\mathbf{r}) = V_{\mathrm{M}}(\mathbf{r}) + V_{\mathrm{R}}(\mathbf{r}), \tag{3.5}$$

the first being the potential generated by ρ_{M} in vacuum and the second being the reaction potential already motivated above. In IEFPCM, these potentials are reformulated as integral equations utilising Green's functions. Moreover, the reaction potential must vanish at infinity, meet the requirement

$$-\nabla^2 V_{\mathrm{R}} = 0 \tag{3.6}$$

inside and outside of the cavity and satisfy the jump condition

$$[V_{\mathrm{R}}] = 0, \tag{3.7}$$

on the cavity surface Γ with $[V] = V_{\mathrm{in}} - V_{\mathrm{out}}$. The last equation expresses the continuity of the potential across the surface. From these conditions, it can be inferred that the reaction potential is of the form[10]

$$V_{\mathrm{R}}(\mathbf{x}) = \int_{\Gamma} \frac{\sigma(\mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|} \mathrm{d}\mathbf{y}, \tag{3.8}$$

where the continuous surface charge σ is a unique solution to

$$A\sigma = -g. \tag{3.9}$$

After a series of transformations and simplifications for isotropic solvents, Tomasi et al. derive reduced forms for the integral operators A and g, resulting in the expression

$$\left[2\pi\left(\frac{\varepsilon+1}{\varepsilon-1}\right) - D\right] S\sigma = -\left(2\pi - D\right) V_{\mathrm{M}}. \tag{3.10}$$

The integral operators appearing in this equation are defined by

$$(S\sigma)(\mathbf{x}) = \int_{\Gamma} G(\mathbf{x}, \mathbf{y})\sigma(\mathbf{y})\mathrm{d}\mathbf{y} \tag{3.11}$$

$$(D\sigma)(\mathbf{x}) = \int_{\Gamma} \left[(\vec{\nabla}_{\mathbf{y}} G(\mathbf{x}, \mathbf{y}))\mathbf{n}(\mathbf{y})\right] \sigma(\mathbf{y})\mathrm{d}\mathbf{y} \tag{3.12}$$

as operators on σ, where $\mathbf{n}(\mathbf{y})$ is the normal vector on the surface Γ at the point $\mathbf{y}$ and $G = \|\mathbf{x}-\mathbf{y}\|^{-1}$ is the Green's function of the operator $-\nabla^2$.[10] Thus, only V_{M} is required to obtain the surface charge.

Construction and discretisation of the cavity surface

So far, σ has been defined as a continuous charge distribution over a surface Γ, which still needs to be properly defined and discretised in order to solve the equations presented here numerically.

Although obtaining the cavity within the solvent continuum can be achieved through numerous methods, including solvent-accessible surfaces, solvent-excluded surfaces, or isosurfaces of the electronic density, it is now common practice to approximate the cavity by a set of interlocking spheres, each centred on an atom of the solute. In this instance, the individual sphere radii are often derived from van der Waals radii.[10,74]

The cavity surface and the ASC equation presented in the last section must then be discretised. The cavity surface can be approximated by a partition into a finite set of plane elements called tesserae, each defined by a representative point in space $\mathbf{s}_k \in \Gamma$, an associated surface area A_k and an orientation defined by the surface normal vector $\mathbf{n}$ at $\mathbf{s}_k$.[74] The solvent reaction potential at a point $\mathbf{x}$ obtained from the continuous surface charge σ (eq. 3.8) is therefore approximated by

$$V_{\mathrm{R}}(\mathbf{x}) \approx \sum_k \frac{\sigma(\mathbf{s}_k) A_k}{\|\mathbf{x} - \mathbf{s}_k\|} = \sum_k \frac{q_k}{\|\mathbf{x} - \mathbf{s}_k\|}, \tag{3.13}$$

where q_k represents an ASC placed at $\mathbf{s}_k$. The surface A_k is chosen in such a way that σ can be considered constant within the tesserae.[10] This allows the continuous surface charge distribution to be expressed as the sum of Coulomb potentials.

The IEFPCM equations for isotropic solvents presented in the last section can thus be rewritten in the matrix vector form

$$\mathbf{q} = -\mathbf{K}\mathbf{V}, \tag{3.14}$$

which represents a set of T coupled equations, one for each tessera. Here the $(T \times T)$ matrix

$$\mathbf{K} = \left\{ \left[2\pi \left(\frac{\varepsilon + 1}{\varepsilon - 1} \right) \mathbf{A}^{-1} - \mathbf{D} \right] \mathbf{S} \right\}^{-1} \left[2\pi \mathbf{A}^{-1} - \mathbf{D} \right] \tag{3.15}$$

encodes information about the geometry of the cavity as well as the dielectric constant of the solvent ε, while $\mathbf{q}$ and $\mathbf{V}$ are column vectors, the former containing the ASC charges and the latter the solute potential at $\mathbf{s}_k$.[10]

The matrix element definitions of $\mathbf{S}$ and $\mathbf{D}$ are implementation specific as discussed in detail in ref. [10]. In the following, only the continuous charge variation of the IEFPCM method and the resulting construction of the diagonal and off-diagonal elements of $\mathbf{S}$ and $\mathbf{D}$ will be presented. It was first briefly introduced in ref. [10] and subsequently discussed in detail by Scalmani and Frisch,[74] and represents the standard PCM method used in Gaussian16.[42] It was motivated by the fact that in previous surface partitioning schemes discontinuities in the energy functional could be observed, thus rendering the

PCM method unreliable in terms of exploring the potential energy surface.[74] Scalmani and Frisch were able to construct a smooth PCM energy surface using a modified surface discretisation scheme by York and Warshel,[75] where the surface charge, now given by

$$\sigma(\mathbf{r}) = \sum_k \frac{q_k}{a_k} \phi_k(\mathbf{r}; \mathbf{s}_k, \zeta_k), \tag{3.16}$$

is expressed in terms of normalised spherical Gaussian basis functions

$$\phi_k(\mathbf{r}; \mathbf{s}_k, \zeta_k) = \left(\frac{\zeta_k^2}{\pi}\right)^{3/2} e^{-\zeta_k^2 \|\mathbf{r}-\mathbf{s}_k\|^2}. \tag{3.17}$$

In addition to the parameters already introduced, each tessera k is described by four additional parameters, namely the fraction of the total surface a_k, the exponent ζ_k of the Gaussian, the self-potential factor f_k and the self-field factor g_k. The reasoning[74] behind the optimal choice of the latter three goes far beyond the scope of this work and can only be briefly summarised here: ζ_k can be fitted to obtain the exact Born ion solvation energy,[10,74] which in turn induces a value for f_k, while g_k can be derived from a sum rule derived from Gauss's law.[74]

In this IEFPCM variant, the operators $\mathbf{S}$ and $\mathbf{D}$ are now expressed using Einstein's notation as

$$\begin{aligned} (\mathbf{S})_{ii} &= \frac{f_i}{a_i} \\ (\mathbf{S})_{ij} &= \langle i|j\rangle \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} (\mathbf{D})_{ii} &= \frac{g_i}{a_i} \\ (\mathbf{D})_{ij} &= -\frac{\partial}{\partial \mathbf{s}_j} \langle i|j\rangle \, \hat{\mathbf{n}}_j. \end{aligned} \tag{3.19}$$

The bra-ket notates an integral which takes the form of a standard two-centre electron repulsion integral and can be simplified to

$$\langle i|j\rangle = \frac{\operatorname{erf}(\zeta'_{ij} s_{ij})}{s_{ij}}, \tag{3.20}$$

with $s_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ and $\zeta'_{ij} = \zeta_i \zeta_j / \sqrt{\zeta_i^2 + \zeta_j^2}$, since the basis of $|i\rangle$ and $|j\rangle$ contains only normalised spherical Gaussians.[74]

3.2.2. Coupling of the electrostatic solvent description and QM

Following the formal definition of the IEFPCM procedure, the apparent surface charges derived from this theory must be incorporated into a quantum mechanical formalism. This is done below for the Hartree-Fock method, but PCM has been implemented for a variety of methods.[10,73]

The Fock matrix of an molecule in a vacuum can be written as

$$\mathbf{F}^0 = \mathbf{h}^0 + \mathbf{G}^0(\mathbf{P}), \tag{3.21}$$

where $\mathbf{h}^0$ and $\mathbf{G}^0$ are the one- and two-electron contributions respectively.[11] The argument of the latter indicates that $\mathbf{G}^0$ depends on the one-electron density matrix $\mathbf{P}$.[10] The corresponding energy functional

$$E^0_{\mathrm{HF}} = \langle \Psi_{\mathrm{HF}} | \hat{H}^0 | \Psi_{\mathrm{HF}} \rangle \,, \tag{3.22}$$

with the standard vacuum Hamiltonian, is written in matrix form as

$$E^0_{\mathrm{HF}} = \mathrm{tr}\, \mathbf{P}\mathbf{h}^0 + \frac{1}{2}\,\mathrm{tr}\, \mathbf{P}\mathbf{G}^0(\mathbf{P}) + V_{\mathrm{nn}}, \tag{3.23}$$

where the last term represents the interaction energy between the nuclei and tr is the trace operator.

To incorporate the ASC as defined in equation 3.14 into this formalism, the Fock matrix is extended by charge interaction terms, giving

$$\mathbf{F} = \mathbf{h}^0 + \mathbf{G}^0(\mathbf{P}) + \mathbf{h}^{\mathrm{R}} + \mathbf{X}^{\mathrm{R}}(\mathbf{P}). \tag{3.24}$$

By analogy with eq. 3.21, $\mathbf{h}^{\mathrm{R}}$ and $\mathbf{X}^{\mathrm{R}}$ represent one- and two-electron terms respectively. The interaction of the former can be expressed in two ways. First, as the matrix $\mathbf{j}^{\mathrm{R}}$, which is defined as the interaction between the potential $V^{\mathrm{e}}_{\mu\nu}$ of the electronic charge distribution and the ASC q^n_k calculated from the potential of the nuclear charges, giving

$$(\mathbf{j}^{\mathrm{R}})_{\mu\nu} = \sum_k V^{\mathrm{e}}_{\mu\nu}(\mathbf{s}_k) q^{\mathrm{n}}_k(\mathbf{s}_k) \tag{3.25}$$

Secondly, the calculation of the interaction can be reversed in the matrix $\mathbf{y}^{\mathrm{R}}$, expressed as

$$(\mathbf{y}^{\mathrm{R}})_{\mu\nu} = \sum_k V^{\mathrm{n}}(\mathbf{s}_k) q^{\mathrm{e}}_{k,\mu\nu}(\mathbf{s}_k), \tag{3.26}$$

where now $q^{\mathrm{e}}_{k,\mu\nu}$ is the ASC calculated from the electronic potential and V^{n} is the potential of the nuclei.[10]

These two matrices are formally identical, so the one-electron term in the ASC formulation of the Fock matrix can be expressed as

$$\mathbf{h}^{\mathrm{R}} = \frac{1}{2}(\mathbf{j}^{\mathrm{R}} + \mathbf{y}^{\mathrm{R}}) \tag{3.27}$$

Finally, the functional for the electronic contribution to the free energy in the Hartree-Fock approximation,

$$G_{\mathrm{el}} = \langle \Psi_{\mathrm{HF}} | \hat{H}^{\mathrm{eff}} | \Psi_{\mathrm{HF}} \rangle - \frac{1}{2} \langle \Psi_{\mathrm{HF}} | \hat{V}^{\mathrm{R}} | \Psi_{\mathrm{HF}} \rangle , \tag{3.28}$$

with $\hat{H}^{\mathrm{eff}} = \hat{H}^0 + \hat{V}^{\mathrm{R}}$[73] can be reformulated as

$$G_{\mathrm{el}} = \mathrm{tr}\, \mathbf{P} \left[\mathbf{h}^0 + \frac{1}{2}(\mathbf{j} + \mathbf{y}) \right] + \frac{1}{2} \mathrm{tr}\, \mathbf{P} \left[\mathbf{G}^0(\mathbf{P}) + \mathbf{X}(\mathbf{P}) \right] + \left[\frac{1}{2} U_{\mathrm{nn}} + V_{\mathrm{nn}} \right] , \tag{3.29}$$

where in the PCM notation U_{nn} is the interaction energy between the nuclei and the ASC.[10,73]

In practice, these equations are solved through a single iteration scheme, allowing for simultaneous optimisation of the ASC and wavefunction in each SCF optimisation cycle. This results in lowered cost of PCM solvation compared to older double iteration schemes that required wavefunction convergence prior to updating the solvent charges.[11,76]

3.3. ONIOM-PCM

ONIOM-PCM was first described by Vreven et al. in 2001[11] and later integrated into the Gaussian software, which is likely to have contributed to the current popularity of this method, as it arguably did for its individual components ONIOM and PCM.

ONIOM-PCM combines two multiscale approximations, as described at the beginning of this chapter: First, the whole system is divided into solvent and solute, the former being approximated by the continuum description of PCM. A second multiscale approximation is then applied to the solute. The integration of the ONIOM method into the PCM formalism must therefore define the interaction of the ASC with the individual sub-calculations and, in turn, the calculation of these charges from the charge density of the sub-calculations.

Vreven et al. defined a total of three ONIOM-PCM schemes that provide a series of approximations to this problem, called ONIOM-PCM/A, /B and /C, as well as a fourth scheme, called ONIOM-PCM/X, which does not fit directly into the hierarchy of the previous methods. The discussion of these models begins with an outline of ONIOM-PCM/A, which can be considered the basic ONIOM-PCM model.[11,12] As in the reference publication [11], the discussion is limited to two-layer ONIOM extrapolations.

3.3.1. ONIOM-PCM/A

Vreven et al. suggest that the charge density for an ONIOM2 calculation can be obtained by the ONIOM extrapolation

$$\rho^{\mathrm{ONIOM}} = \Omega(\boldsymbol{\rho}). \tag{3.30}$$

This is an extension of the ONIOM energy calculation philosophy to the calculation of other properties. The authors therefore refer to the ONIOM-PCM/A scheme as the "correct" method,[11] although it should be noted that this is still an ad hoc approximation to the *hr*-density.

Using this ansatz, the ASC defined by eq. 3.14 can be rewritten as

$$\mathbf{q}^{\mathrm{ONIOM}} = -\mathbf{K}\mathbf{V}^{\mathrm{ONIOM}} = -\mathbf{K}(\mathbf{V}^{\mathrm{lr}} + \mathbf{V}^{\mathrm{hm}} - \mathbf{V}^{\mathrm{lm}}), \tag{3.31}$$

where the contributions to the potential vector $\mathbf{V}^{\mathrm{ONIOM}}$ are obtained from the individual charge densities in eq. 3.30. The cavity is constructed in the same way as in the standard PCM or in the ONIOM notation *hr*-PCM method, taking into account the geometry of the *real* system. This cavity is then used for all sub-calculations, i.e. the calculations on the *model* system are also performed in the *lr*-cavity. The solute potentials at the characteristic points of the cavity are obtained by an ONIOM extrapolation of the potentials as indicated in the last equation.

The electronic contribution to the free energy is therefore written as

$$G_{\mathrm{el}}^{/\mathrm{A}} = G_{\mathrm{el}}^{\mathrm{lr}}(q^{\mathrm{ONIOM}}) + G_{\mathrm{el}}^{\mathrm{hm}}(q^{\mathrm{ONIOM}}) - G_{\mathrm{el}}^{\mathrm{lm}}(q^{\mathrm{ONIOM}}) \tag{3.32}$$

The arguments to the free energies indicate that they are evaluated using the ASC obtained from the ONIOM potential $\mathbf{V}^{\mathrm{ONIOM}}$.[11] Since the ASCs depend on the result of all three sub-calculations, ONIOM-PCM/A has been implemented using a double iteration scheme. This means that, starting from a first estimate of the solute potential from the vacuum or another suitable method, all sub-calculations are converged with the set of ONIOM-ASCs. A new set of ASCs is then calculated from the updated solute potential and the process is repeated. Thus, the ASC calculation and the convergence of the sub-calculations are nested in an additional iteration loop, hence the name double iteration scheme. This is computationally less efficient than the single iteration scheme for the standard IEFPCM procedure, where the ASC is recalculated at each SCF step.[11] Although Vreven et al. state that it is theoretically possible to construct such a single iteration scheme for ONIOM-PCM/A, they stress that it would require a significant restructuring of their code. This is in contrast to the limited amount of changes that are usually required to use the ONIOM method.[11] The other ONIOM-PCM schemes were therefore motivated by the intention to develop alternatives that would allow the use of the standard IEFPCM single iteration approach.

3.3.2. ONIOM-PCM/B

The second scheme ONIOM-PCM/B uses the same cavity as the /A scheme, but approximates the ONIOM charge density by the *lr* charge density. The calculation of the ASC is thus decoupled from the two *model* system calculations and it is possible to first perform a standard PCM calculation at the *low* level of theory and then calculate the

model system energies in the fixed field of the solvent charges. The ASC are calculated with the *lr* solute potential by

$$\mathbf{q}^{\mathrm{lr}} = -\mathbf{K}\mathbf{V}^{\mathrm{lr}}. \tag{3.33}$$

As a consequence, the free energy for the /B scheme is formulated as

$$G_{\mathrm{el}}^{/\mathrm{B}} = G_{\mathrm{el}}^{\mathrm{lr}}(q^{\mathrm{lr}}) + E^{\mathrm{hm}}(q^{\mathrm{lr}}) - E^{\mathrm{lm}}(q^{\mathrm{lr}}). \tag{3.34}$$

The central approximation of this approach, the sole use of the *lr* potential to calculate the solvent charges, is justified by the authors by the expectation that the contributions of the *model* system potentials at the cavity surface largely cancel. Implicitly,

$$\mathbf{V}^{\mathrm{hm}} - \mathbf{V}^{\mathrm{lm}} = \mathbf{0} \tag{3.35}$$

applies and the potentials are set equal. As a result, ONIOM-PCM/B can be seen as an approximation to /A.

3.3.3. ONIOM-PCM/C

In addition to the approximations introduced in the ONIOM-PCM/B scheme, /C evaluates the *model* system calculations in vacuum, thus excluding the solute-solvent interaction. The simplified free energy expression is then

$$G_{\mathrm{el}}^{/\mathrm{C}} = G_{\mathrm{el}}^{\mathrm{lr}}(q^{\mathrm{lr}}) + E_{\mathrm{vac}}^{\mathrm{hm}} - E_{\mathrm{vac}}^{\mathrm{lm}}. \tag{3.36}$$

This is a drastic approximation, assuming not only that the solute *model* system potentials cancel, but also that the extrapolation to *hr* is not affected by the solvent field.[12] In other words, it is assumed that the energy difference between the *model* systems in the solvent charge field is equal to the difference in vacuum

$$E^{\mathrm{hm}}(q^{\mathrm{lr}}) - E^{\mathrm{lm}}(q^{\mathrm{lr}}) = E_{\mathrm{vac}}^{\mathrm{hm}} - E_{\mathrm{vac}}^{\mathrm{lm}}. \tag{3.37}$$

By separating the energies of *hm* and *lm*, the last statement can be reformulated as

$$E^{\mathrm{hm}}(q^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{hm}} =: \Delta_{\mathrm{pol}}E^{\mathrm{hm}} = \Delta_{\mathrm{pol}}E^{\mathrm{lm}} := E^{\mathrm{lm}}(q^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{lm}}, \tag{3.38}$$

which is equivalent to saying that /C leads to identical results as /B, if the *low* and *high* levels of theory give identical results for the polarisation energy induced by the *lr* solvent charges on the *lr* cavity.

The last three ONIOM-PCM schemes can thus be ordered hierarchically by increasing level of approximation. Starting from /A, where the *model* system charge density polarises the solvent and the solvent can polarise the *model* system in both calculations, /B is obtained by neglecting the effect of the *model* system charge density on the solvent. In turn, /C introduces an additional approximation by also neglecting the opposite

direction of mutual polarisation, i.e. the influence of the solvent on the *model* systems. Although it should be emphasised that these approximations only neglect the polarisation effect on the individual *model* system calculation, they do not completely neglect the polarisation effect on the set of *model* system atoms, which by definition are part of the *real* system. These effects are in all cases modelled by the *lr* calculation and schemes /B and /C only decouple the polarisation from the *high* level theory and therefore from the ONIOM extrapolation. In both cases the basic assumption is that the contributions from the *model* system are cancelled out by the intrinsic error correction of the ONIOM method, which is achieved by the difference between the *hm* and *lm* calculations in the ONIOM extrapolation.

3.3.4. ONIOM-PCM/X

The free energy of the last ONIOM-PCM method is obtained by performing three separate PCM calculations and finally an ONIOM extrapolation on the resulting free energies. The three sets of ASC required for this are obtained directly from the solute charge density of the respective sub-calculations, i.e. each calculation is performed with its own set of solvent charges.[11,12]. As with all ONIOM-PCM variants, the *lr* cavity is used for all sub-calculations. The expression for the free energy is therefore

$$G_{\mathrm{el}}^{/\mathrm{X}} = G_{\mathrm{el}}^{\mathrm{lr}}(q^{\mathrm{lr}}) + G_{\mathrm{el}}^{\mathrm{hm}}(q^{\mathrm{hm}}) - G_{\mathrm{el}}^{\mathrm{lm}}(q^{\mathrm{lm}}). \tag{3.39}$$

Unlike /B and /C, this approximation cannot be regarded as an approximation to /A, and the developers of the ONIOM-PCM family of methods therefore state that it must be regarded as an alternative to /A.[11]

Furthermore, Vreven et al. consider /X to be less consistent with ONIOM because of the use of multiple "reaction fields".[11] However, it is worth noting that ONIOM cannot be derived from first principles and therefore should be considered an ad hoc approximation. It is therefore just as reasonable to postpone the ONIOM step in the free energy approximation by performing it after the PCM calculations, as in /X, as to perform the ONIOM extrapolation on the solute charge densities first, as in /A. Following this argument, /A and /X share the same level of approximation and neither can be considered more in line with the ONIOM method than the other.

4. EC-RISM and its ONIOM approximation

The central model of this work is the ONIOM extension of the EC-RISM solvation model. The final steps required to describe this model are therefore to present the theory of the EC-RISM method on which it is based.

As in PCM, the general idea behind EC-RISM is to polarise the solute by a set of point charges representing the solvent. However, the solvent is not modelled as a continuum, but by distribution functions obtained from the solute-solvent interaction potential. This allows a granular and therefore more refined description of the solvent. The resulting solvent charge distribution is mapped onto point charges, which are used to polarise the solute. In turn, the electrostatic potential of the polarised solute can influence the solvent distribution, allowing mutual solute-solute polarisation.

This general approach is similar to the PCM methodology. It is therefore useful to outline the theory in a similar way to the last chapter: First the calculation of the solvent structure and the associated models (1D-RISM, 3D-RISM) and then the coupling of the solvent description with the quantum mechanical description of the solute (EC-RISM). Finally, the ONIOM extension to EC-RISM is presented. We start with the definition of distribution and correlation functions, the central quantities of statistical solvation models.

4.1. Solvent distribution and correlation functions

The microscopic structure of an atomic liquid consisting of N particles can be expressed by its single particle density[77]

$$\rho_N^{(1)}(\mathbf{r}) = \left\langle \sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{r}_i) \right\rangle, \tag{4.1}$$

by averaging over single configurations, defined through the set of all spatial coordinates $\mathbf{r}_i$ of the atoms i. In a similar way the two particle density can be obtained by

$$\rho_N^{(2)}(\mathbf{r}, \mathbf{r}') = \left\langle \sum_{i=1}^{N} \sum_{j=1}^{N}{}' \delta(\mathbf{r} - \mathbf{r}_i)\delta(\mathbf{r}' - \mathbf{r}_j) \right\rangle, \tag{4.2}$$

omitting summation over terms with $i = j$, as indicated by the prime on the second summation sign.[77]

Using these two functions, the pair distribution function is formally defined as

$$g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \frac{\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2)}{\rho_N^{(1)}(\mathbf{r}_1)\rho_N^{(1)}(\mathbf{r}_2)}. \tag{4.3}$$

With the generalised n-particle density function $\rho_N^{(n)}(\mathbf{r}^n)$, the pair distribution function can be also generalised to the n-particle distribution function

$$g_N^{(n)}(\mathbf{r}^n) = \frac{\rho_N^{(n)}(\mathbf{r}_1, \ldots, \mathbf{r}_n)}{\prod_{i=1}^{n} \rho_N^{(1)}(\mathbf{r}_i)}. \tag{4.4}$$

At infinite separation of the n particles, the function approaches $1 - 1/N$ in the canonical ensemble.[77] Hansen and McDonald, in their book on fluid theory, state that the particle distribution function measures the extent to which the structure of a fluid deviates from complete homogeneity.[77]

In addition to the structural insight provided by the pair distribution function, it can also be used to derive thermodynamic properties. For example, the excess internal energy for a simple atomic liquid in the canonical ensemble, where the individual particles interact through pairwise additive forces, can be expressed as the integral[77]

$$\frac{U^{\mathrm{ex}}}{N} = 2\pi\rho \int_0^\infty v(r)g(r)r^2 \,\mathrm{d}r. \tag{4.5}$$

Here $g(r)$ represents the radial distribution function for isotropic systems with the argument representing the separation $r = \|\mathbf{r}_2 - \mathbf{r}_1\|$, while $v(r)$ is the pair potential of the system.

Although the summation over delta functions in equations 4.1 and 4.2 give a rather intuitive definition of the particle density and the resulting distribution functions, they require the ensemble average of the particle coordinates. Molecular dynamics simulations can be used to obtain these quantities, but require extensive sampling of the configurational space of the fluid. To avoid this, statistical solvation models, such as the "reference interaction site model" (RISM) family of methods, use classical density functional theory to derive the equilibrium distribution functions. References [77], [78] and [79] provide an extensive, although challenging overview of the topic and will be used here to briefly summarise the theory behind the approach.

It will be useful for the discussion to work in the grand canonical ensemble. By dropping the subscript N, equations 4.1, 4.2 and 4.4 also hold in this ensemble, and the pair distribution function $g^2(\mathbf{r}_1, \mathbf{r}_2)$ now approaches 1 as the separation of the two particles approaches infinity, i.e. $\|\mathbf{r}_2 - \mathbf{r}_1\| \to \infty$, which avoids some problems during integration.[77] At the same time, the total pair correlation function, another central quantity defined as

$$h^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = g^{(2)}(\mathbf{r}_1, \mathbf{r}_2) - 1 \tag{4.6}$$

approaches 0 at infinite separation.

Another advantage of working in the grand canonical ensemble is that the particle densities defined above through sums over delta functions can be expressed as functional derivatives of the grand partition function. For a system of identical spherical particles in an external field $\phi(\mathbf{r})$, the partition function can be defined as

$$\Xi = \sum_{N=0}^{\infty} \frac{1}{N!} \int \cdots \int \exp\left(-\beta V_N\right) \left(\prod_{i=1}^{N} z^*(i) \right) \mathrm{d}1 \ldots \mathrm{d}N, \tag{4.7}$$

where $\mathrm{d}1 \ldots \mathrm{d}N$ denotes integration over the coordinates $(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ and V_N is the inter-particle potential energy.[77] The local activity

$$z^*(\mathbf{r}) = \frac{\exp\left(\beta\psi(\mathbf{r})\right)}{\Lambda^3} \tag{4.8}$$

can be calculated using the thermal wavelength Λ and the intrinsic chemical potential

$$\psi(\mathbf{r}) = \mu - \phi(\mathbf{r}). \tag{4.9}$$

The grand partition function can then be used with its functional derivative with respect to the local activity

$$\frac{\delta\Xi}{\delta z^*(1)} = \sum_{N=1}^{\infty} \frac{1}{(N-1)!} \int \cdots \int \exp\left(-\beta V_N\right) \left(\prod_{i=2}^{N} z^*(i) \right) \mathrm{d}2 \ldots \mathrm{d}N \tag{4.10}$$

as a generating functional for the n-particle density[77–79]

$$\rho^{(n)}(1, \ldots, n) = \frac{z^*(1) \ldots z^*(n)}{\Xi} \frac{\delta^n \Xi}{\delta z^*(1) \ldots \delta z^*(n)}. \tag{4.11}$$

4.2. The Ornstein-Zernike equation

The previous steps can be used to establish the primary equation of statistical solvation models, the Ornstein-Zernike equation, which links the total pair correlation function $h^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$ to the direct pair correlation function $c^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$, which will be introduced later. In the following, the superscript indicating their definition as two-particle functions will be omitted for brevity and will only be shown explicitly when necessary.

From the theory of functional derivatives, we know that

$$\frac{\delta z^*(\mathbf{r}_1)}{\delta z^*(\mathbf{r}_2)} = \delta(\mathbf{r}_1 - \mathbf{r}_2). \tag{4.12}$$

Upon inserting this equation into the functional derivatives' chain rule equivalent, we get

$$\int \frac{\delta\rho^{(1)}(\mathbf{r}_1)}{\delta \ln z^*(\mathbf{r}_3)} \frac{\delta \ln z^*(\mathbf{r}_3)}{\delta\rho^{(1)}(\mathbf{r}_2)} \,\mathrm{d}\mathbf{r}_3 = \delta(\mathbf{r}_1 - \mathbf{r}_2). \tag{4.13}$$

The two terms in the product are given by

$$\frac{\delta\rho^{(1)}(\mathbf{r}_1)}{\delta \ln z^*(\mathbf{r}_2)} = \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2) - \rho^{(1)}(\mathbf{r}_1)\rho^{(1)}(\mathbf{r}_2) + \delta(\mathbf{r}_1 - \mathbf{r}_2)\rho^{(1)}(\mathbf{r}_1) \tag{4.14}$$

and

$$\frac{\delta \ln z^*(\mathbf{r}_1)}{\delta\rho^{(1)}(\mathbf{r}_2)} = \frac{\delta(\mathbf{r}_1 - \mathbf{r}_2)}{\rho^{(1)}(\mathbf{r}_2)} - c(\mathbf{r}_1, \mathbf{r}_2), \tag{4.15}$$

respectively.[78] The latter equation provides the first formal definition of the direct pair correlation function. By inserting equations 4.14 and 4.15 into equation 4.13 and performing some manipulations,[78,79] the Ornstein-Zernike (OZ) equation

$$h(\mathbf{r}_1, \mathbf{r}_2) = c(\mathbf{r}_1, \mathbf{r}_2) + \int c(\mathbf{r}_1, \mathbf{r}_3)\rho(\mathbf{r}_3)h(\mathbf{r}_3, \mathbf{r}_2) \,\mathrm{d}\mathbf{r}_3 \tag{4.16}$$

can be derived. Since the total correlation function appears not only on the left-hand side of the equation but also under the integral sign, it can be rephrased as the recursive expansion

$$\begin{aligned} h(1,2) = c(1,2) &+ \int c(1,3)\rho^{(1)}(3)c(3,2) \,\mathrm{d}3 \\ &+ \iint c(1,3)\rho^{(1)}(3)c(3,4)\rho^{(1)}(4)c(4,2) \,\mathrm{d}3\mathrm{d}4 + \ldots, \end{aligned} \tag{4.17}$$

again using the shorthand notation $1, \ldots, N$ for particle coordinates $\mathbf{r}_1, \ldots, \mathbf{r}_N$.[77] From this representation it can be seen that the Ornstein-Zernike equation not only establishes a link between the total and direct correlation functions, but also provides a physical interpretation for both: The total correlation function between particles 1 and 2 is partitioned into a direct correlation between those particles $c(1,2)$ and an indirect correlation that is defined through integrals over c and $\rho^{(1)}$. Equation 4.17 shows that this indirect part is evoked by the propagation of the correlation through the other particles present in the system. The recursion thus encodes the indirect interaction of particles 1 and 2 through an increasing number of additional particles.[77]

To solve the Ornstein-Zernike equation, an additional closure relation between h and c is needed to eliminate one of the unknown functions. This closure generally follows the form[78]

$$c(\mathbf{r}_1, \mathbf{r}_2) = \exp\left[-\beta u(\mathbf{r}_1, \mathbf{r}_2) + t(\mathbf{r}_1, \mathbf{r}_2) + b(\mathbf{r}_1, \mathbf{r}_2)\right] - 1 - t(\mathbf{r}_1, \mathbf{r}_2), \tag{4.18}$$

where $u(\mathbf{r}_1, \mathbf{r}_2)$ is the interaction potential between a pair of particles and

$$t(\mathbf{r}_1, \mathbf{r}_2) = h(\mathbf{r}_1, \mathbf{r}_2) - c(\mathbf{r}_1, \mathbf{r}_2). \tag{4.19}$$

The bridge function $b(\mathbf{r}_1, \mathbf{r}_2)$ entails many-body integrals with a solution that poses a significant challenge and requires numerical approximations.[79] One possible approach to tackle this issue is to neglect this bridge function, which gives the Hypernetted Chain Closure (HNC)

$$c(\mathbf{r}_1, \mathbf{r}_2) = \exp\left[-\beta u(\mathbf{r}_1, \mathbf{r}_2) + t(\mathbf{r}_1, \mathbf{r}_2)\right] - 1 - t(\mathbf{r}_1, \mathbf{r}_2). \tag{4.20}$$

The full, albeit challenging, mathematical derivation of the HNC closure can be found in references [80], [81] or [77], among others. The first also provides a brief historical overview of its development.

In brief, the HNC closure can be formally derived by a cluster expansion of the total correlation function h in the densities $\rho^{(1)}$. The resulting series becomes increasingly notationally complex with increasing order of the expansion and is therefore more conveniently expressed as a graph representation of the expansion terms, called cluster diagrams.[80] The exact solution of h is found by summation over the set of all cluster diagrams, which finally results in equation 4.18. Approximate solutions are therefore found by omitting cluster diagrams from the summation.[80] The interactions encoded in these diagrams represent the direct and indirect interactions of the particles of the fluid and are conceptually similar to the expansion of the Ornstein-Zernike equation shown in equation 4.17, in fact it can also be derived in the form of a diagrammatic expansion.[77]

The complete set of cluster diagrams that occur in the cluster expansion of h is generally partitioned into subsets based on their topology. Different partitions are used in the literature. Rowlinson, in his review of the equation of states of dense systems,[80] uses a partition of cluster diagrams into four classes called simple chains, netted chains, bundles and elementary clusters, the second of which gives rise to the name of the HNC closure. Klein and Green identify three classes of cluster diagrams, called series, parallel and bridge diagrams.[81] However, in more recent publications, such as [82] and [77], it is more common to use only a classification into series and bridge diagrams. The class of bridge diagrams gives rise to the term b in equation 4.18. The approximation of the HNC closure omits all bridge diagrams from the cluster expansion, which is identical to setting $b = 0$, giving equation 4.20. Other alternative closures will be presented throughout this chapter.

The equations presented thus far describe systems made up of identical, spherical particles. To develop a theory capable of describing molecular systems, it is necessary to extend the Ornstein-Zernike equation. By considering the orientation of interacting molecules, the Molecular Ornstein-Zernike equation can be expressed as

$$h(1,2) = c(1,2) + \left(\frac{\rho}{\Omega_a}\right) \int c(1,3)h(3,2)\ \mathrm{d}3. \tag{4.21}$$

Here, the function arguments $(i, j) = (\{\mathbf{r}_i, \Omega_i\}, \{\mathbf{r}_j, \Omega_j\})$ represent not only the spatial but also angular coordinates, indicating the molecular orientation. Hence, the integration takes place over all these coordinates and (ρ/Ω_a) represents a normalisation constant, with the mean particle density ρ and the solid angle Ω_a.[79] For this equation, a solution is available that is based on spherical harmonic expansions.[83] However, it is considered to have slow convergence for non-spherical systems.[78,79] The following theories thus aim to provide an approximate solution to the Molecular Ornstein-Zernike equation that is applicable to more complex geometries.

4.3. 1D-RISM

The "reference interaction site model" (RISM), first introduced by Andersen and Chandler in 1972[84,85] and here named 1D-RISM to distinguish it from the later introduced three-dimensional case, approximates the Molecular Ornstein-Zernike equation by decomposing the interaction between two molecules into site-site interactions. These reference sites can be atoms of the molecule or other sites, which enable a representative description of the molecular interaction. The interaction potential between two molecules is thus expressed as

$$u(1,2) = \sum_{\alpha} \sum_{\gamma} u_{\alpha\gamma}(r), \tag{4.22}$$

where α and γ denote the reference sites.[78,79] Hence, this function solely depends on the inter-site distance, $r = \|\mathbf{r}_1 - \mathbf{r}_2\|$. In addition, the second fundamental approximation of RISM is that the direct pair correlation function is approximated in a similar way as a sum of site-site direct correlation functions

$$c(1,2) = \sum_{\alpha} \sum_{\gamma} c_{\alpha\gamma}(r). \tag{4.23}$$

This second approximation allows the averaging over the angular coordinates to be performed analytically in the Molecular Ornstein-Zernike equation, which, after some further manipulation,[78,79] gives the matrix form of the RISM equation

$$\mathbf{h} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\rho\omega} * \mathbf{c} * \mathbf{h}. \tag{4.24}$$

Here $\cdot * \cdot$ denotes the matrix product and convolution of two matrices, $(\mathbf{h})_{\alpha\gamma} = h_{\alpha\gamma}(r)$ and $(\mathbf{c})_{\alpha\gamma} = c_{\alpha\gamma}(r)$. The matrix $\boldsymbol{\omega}$ is defined as

$$(\boldsymbol{\omega})_{\alpha\gamma} = \omega_{\alpha\gamma}(r) = \delta_{\alpha\gamma}\delta(r) + (1 - \delta_{\alpha\gamma})s_{\alpha\gamma}(r) \tag{4.25}$$

with the intramolecular correlation function

$$s_{\alpha\gamma}(r) = \frac{1}{4\pi L_{\alpha\gamma}^2}\delta(r - L_{\alpha\gamma}). \tag{4.26}$$

This function encodes a rigid molecular geometry via the distance $L_{\alpha\gamma}$ between the α and γ sites.[78] The basic RISM model thus represents a one-dimensional approximation to the formally six-dimensional Molecular Orstein-Zernike equation by averaging the orientations of the molecules at a fixed site separation r.[86]

The 1D-RISM equations presented so far describe a liquid system containing only one molecular species. In the early 1980s, Hirata and collaborators extended the theory to describe mixtures by, among other things, introducing a renormalisation scheme that solves the problem of divergent integrals of the electrostatic interaction terms in $u(r)$. The resulting model was named the "Extended Reference Interaction Site Model" (XRISM),[87–89] which will be also referred to as 1D-RISM for simplicity. After applying the concept of infinite dilution the outcome for solute-solvent systems is the set of equations

$$\mathbf{h}^{\mathrm{vv}} = \mathbf{w}^{\mathrm{v}} * \mathbf{c}^{\mathrm{vv}} * \mathbf{w}^{\mathrm{v}} + \mathbf{w}^{\mathrm{v}} * \mathbf{c}^{\mathrm{vv}} * \boldsymbol{\rho}^{\mathrm{v}}\mathbf{h}^{\mathrm{vv}} \tag{4.27}$$

$$\mathbf{h}^{\mathrm{uv}} = \mathbf{w}^{\mathrm{u}} * \mathbf{c}^{\mathrm{uv}} * \mathbf{w}^{\mathrm{v}} + \mathbf{w}^{\mathrm{u}} * \mathbf{c}^{\mathrm{uv}} * \boldsymbol{\rho}^{\mathrm{v}}\mathbf{h}^{\mathrm{vv}} \tag{4.28}$$

$$\mathbf{h}^{\mathrm{uu}} = \mathbf{w}^{\mathrm{u}} * \mathbf{c}^{\mathrm{uu}} * \mathbf{w}^{\mathrm{u}} + \mathbf{w}^{\mathrm{u}} * \mathbf{c}^{\mathrm{uv}} * \boldsymbol{\rho}^{\mathrm{v}}\mathbf{h}^{\mathrm{vu}}, \tag{4.29}$$

with $\mathbf{w}^{v} = (\boldsymbol{\rho}^{\mathrm{v}})^{-1}\boldsymbol{\omega}^{\mathrm{v}}$ and $\mathbf{w}^{\mathrm{u}} = (\boldsymbol{\rho}^{\mathrm{u}})^{-1}\boldsymbol{\omega}^{\mathrm{u}}$, describing the correlation functions for the three interaction modes: Solvent-solvent (vv), solute-solvent (uv) and solute-solute (uu).[78,79]

The first two equations can be shortened to

$$\boldsymbol{\rho}^{\mathrm{v}}\mathbf{h}^{\mathrm{vv}} = \boldsymbol{\omega}^{\mathrm{v}} * \mathbf{c}^{\mathrm{vv}} * \boldsymbol{\chi}^{\mathrm{vv}} \tag{4.30}$$

$$\boldsymbol{\rho}^{\mathrm{v}}\mathbf{h}^{\mathrm{uv}} = \boldsymbol{\omega}^{\mathrm{u}} * \mathbf{c}^{\mathrm{uv}} * \boldsymbol{\chi}^{\mathrm{vv}}, \tag{4.31}$$

using the definition for the site-site solvent susceptibility

$$\boldsymbol{\chi}^{\mathrm{vv}} = \boldsymbol{\rho}^{\mathrm{v}}\boldsymbol{\omega}^{\mathrm{v}} + (\boldsymbol{\rho}^{\mathrm{v}})^{2}\mathbf{h}^{\mathrm{vv}}. \tag{4.32}$$

In practice, precomputed solvent susceptibilities are used to solve the RISM equations for solute-solvent systems, which are obtained from the "Dielectrically Consistent Reference Interaction Site Model" (DRISM),[90–92] a variation of the 1D-RISM approach, that applies an effective bridge function to yield improved, dielectric consistent results for finite-concentration ionic solutions.

4.4. 3D-RISM

The one-dimensional RISM approach represents a drastic reduction of the complex molecular Ornstein-Zernike equation to the radial symmetric case. This results in a loss of information about the true three-dimensional solvent distribution.[6] To recover some of the microscopic structure of the liquid system, multiple contributors generalised

1D-RISM to the three-dimensional case, consequently called 3D-RISM.[6–8] As in the earlier model, it is derived from the Molecular Ornstein-Zernike equation by decomposing the molecular interactions into site-site interactions and approximating the direct solute-solvent correlation function as the sum of the site-site direct correlation functions. In contrast to the one-dimensional case, this includes averaging over the solvent orientations, but not over the solute coordinates.[86] The result is the central equation[3]

$$\rho_\gamma h_\gamma(\mathbf{r}) = \sum_{\gamma'} c_{\gamma'} * \chi_{\gamma\gamma'}(\mathbf{r}), \tag{4.33}$$

where γ and γ' are solvent sites, ρ_γ is the bulk density of the solvent and $\chi_{\gamma\gamma'}$ is the site-site solvent susceptibility, as defined in equation 4.32. As only the solute-solvent case is relevant for further discussion, the superscripts have been omitted.

Similar to equation 4.18, the HNC closure in the context of 3D-RISM is defined as

$$h_\gamma(\mathbf{r}) = \exp\left[t^*_\gamma(\mathbf{r})\right] - 1 \tag{4.34}$$

with

$$t^*_\gamma(\mathbf{r}) = -\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}). \tag{4.35}$$

It is often observed that numerical solutions to the 3D-RISM equations using the HNC closure do not converge due to their highly non-linear nature.[3] In an effort to solve this problem, Kovalenko and Hirata presented an ad hoc,[79] partially linearised[3] modification called the KH-closure,[86,93] defined by

$$h_\gamma(\mathbf{r}) = \begin{cases} \exp\left[t^*_\gamma(\mathbf{r})\right] - 1, & \text{if } t^*_\gamma(\mathbf{r}) \leq 0 \\ t^*_\gamma(\mathbf{r}), & \text{otherwise.} \end{cases} \tag{4.36}$$

In 2003, Kast showed the conditions that must be satisfied for non-zero bridge functions, i.e. closures beyond the HNC approximation, to be path independent.[94,95] Building on this, in 2008, Kast and Kloss presented another closure called the partial series expansion of order n (PSE-n), which encompasses the last two closures.[96] It is expressed as

$$h_\gamma(\mathbf{r}) = \begin{cases} \exp\left[t^*_\gamma(\mathbf{r})\right] - 1, & \text{if } t^*_\gamma(\mathbf{r}) \leq 0 \\ \displaystyle\sum_{i=0}^{n} \left[\frac{(t^*_\gamma(\mathbf{r}))^i}{i!}\right] - 1, & \text{otherwise} \end{cases} \tag{4.37}$$

and is identical to the KH closure for $n = 1$ and gives the HNC closure for the limit $n \to \infty$. The authors found that the PSE-n closure is stable for small n, while still giving similar results to the HNC closure.[96]

The basic idea of scaling between KH and HNC closures is mathematically simple. The KH closure divides the codomain of t^*_γ into two cases and only the case for $t^*_\gamma > 0$

differs from HNC. The function $\exp[t^*_\gamma]$ is now expressed as its power series expansion as shown in eq. 4.37. The partial linearisation applied by Kovalenko and Hirata for $t^*_\gamma > 0$ represent the terms up to $i = 1$ of this sum. The limit $n \to \infty$ therefore recovers the exponential function and the unpartitioned codomain of t^*_γ as used in the HNC closure.

From the correlation functions presented here it is possible to calculate a variety of thermodynamic properties of the fluid system. In the context of EC-RISM it will be important to calculate the excess chemical potential for reasons that will be explained later. With the HNC closure the excess chemical potential can be calculated via the expression

$$\mu_{\mathrm{ex}}^{\mathrm{HNC}} = \beta^{-1} \sum_\gamma \rho_\gamma \int \frac{1}{2} h_\gamma^2(\mathbf{r}) - c_\gamma(\mathbf{r}) - \frac{1}{2} h_\gamma(\mathbf{r}) c_\gamma(\mathbf{r}) \, \mathrm{d}\mathbf{r}. \tag{4.38}$$

For the PSE-n closure, a second term is added that depends on the order of the series expansion

$$\mu_{\mathrm{ex}}^{\text{PSE-}n} = \mu_{\mathrm{ex}}^{\mathrm{HNC}} - \beta^{-1} \sum_\gamma \rho_\gamma \int \frac{\Theta(h_\gamma(\mathbf{r}))(t^*_\gamma(\mathbf{r}))^{n+1}}{(n+1)!} \, \mathrm{d}\mathbf{r}, \tag{4.39}$$

which vanishes for $n \to \infty$, giving the HNC result.[96] Here Θ is the Heaviside step function. The numerical considerations required to solve the 3D-RISM equations and compute the excess chemical potential can be found in references [97] and [17].

Another important thermodynamic quantity that will be used later to empirically correct the excess chemical potential is the partial molar volume (PMV), which is formally defined as the partial derivative of the volume with respect to the amount of substance i

$$V_{\mathrm{m}} = \left(\frac{\partial V}{\partial n_i} \right)_{T,p,n_{i \neq j}}. \tag{4.40}$$

From Kirkwood-Buff theory combined with 3D-RISM, the dimensionless PMV ρV_{m} of the solute at infinite dilution can be expressed as

$$\rho V_{\mathrm{m}} = \rho \beta^{-1} \chi_T \left(1 - \rho \sum_\gamma \int c_\gamma(\mathbf{r}) \, \mathrm{d}\mathbf{r} \right), \tag{4.41}$$

where χ_T and ρ are the isothermal compressibility and density of the pure solvent, respectively.[98–100]

The last part of the basic 3D-RISM model to be discussed is the solute-solvent-site interaction potential u_γ. The choice of model potential depends on the scale of the sites selected in the RISM model. Usually an atomistic scale is chosen, so the sites are located at the atomic coordinates $\mathbf{R}_\alpha$, and a simple force field term for non-bonded interactions, in most cases the sum of a Lennard-Jones and an electrostatic term

$$u_\gamma(\mathbf{r}) = u_\gamma^{\mathrm{LJ}}(\mathbf{r}) + u_\gamma^{\mathrm{el}}(\mathbf{r}), \tag{4.42}$$

with

$$u_\gamma^{\mathrm{LJ}}(\mathbf{r}) = \sum_\alpha 4\varepsilon_{\alpha\gamma}\left[\left(\frac{\sigma_{\alpha\gamma}}{\|\mathbf{r}-\mathbf{R}_\alpha\|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{\|\mathbf{r}-\mathbf{R}_\alpha\|}\right)^{6}\right] \tag{4.43}$$

and the electrostatic term being approximated by an Coulomb potential

$$u_\gamma^{\mathrm{el}}(\mathbf{r}) \approx u_\gamma^{\mathrm{C}}(\mathbf{r}) = q_\gamma \sum_\alpha \frac{q_\alpha}{\|\mathbf{r}-\mathbf{R}_\alpha\|}. \tag{4.44}$$

This allows for the polarisation of the solvent by interaction with the solute. Mutual polarisation is only possible if the solute is described at a level that allows polarisation of the solute by the solvent. A way to achieve this will be presented in the following section.

4.5. EC-RISM

To allow for mutual polarisation, Kloss et al. presented an extension to the 3D-RISM model, called "embedded cluster reference interaction site model" (EC-RISM),[3] which couples the statistical solvation model with a quantum mechanical description of the solute. This model represents an alternative formulation to previous 3D-RISM based SCF-models and a brief comparison is given in section 4.5.2.

In EC-RISM the free energy of the solute is approximated as the sum of the quantum mechanical energy of the polarised solute and its excess chemical potential

$$G_{\mathrm{sol}} = E_{\mathrm{sol}}(q_{\mathrm{solv}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}, u_{\mathrm{LJ}}). \tag{4.45}$$

The former depends on the set of solvent charges q_{solv} derived from the solvent partition functions. The latter depends on the solute potential φ_{sol} and the Lennard-Jones-Potential u_{LJ} and is evaluated using 3D-RISM. In order to outline EC-RISM, the solute-solvent interaction potential has to be constructed similar to eq. 4.42 based on the quantum mechanical description of the solute. Furthermore, a way has to be found to incorporate the polarising effect of the solute distribution into the solute Hamiltonian. We start with the first problem.

4.5.1. Coupling the statistical solvent description with QM

As in the standard 3D-RISM approach, the solute-solvent interaction is split into a Lennard-Jones and an electrostatic term. The latter can be calculated directly from the solute electrostatic potential (ESP) φ_{sol} with

$$u_\gamma^{\mathrm{el}}(\mathbf{r}) = q_\gamma \varphi_{\mathrm{sol}}(\mathbf{r}). \tag{4.46}$$

In the original formulation of the EC-RISM model, the ESP was approximated by fitted atomic partial charges, giving an approximation as in eq. 4.44.[3] This model is referred

to as the "point charge model", while the use of the "exact" potential φ_{sol} is usually referred to as the "full potential model" or "exact potential model".

The solvent distribution g_γ derived from this potential via 3D-RISM can then be converted with the solvent site density ρ_γ and charge q_γ into the charge density

$$\rho_q(\mathbf{r}) = \sum_\gamma q_\gamma \rho_\gamma g_\gamma(\mathbf{r}), \tag{4.47}$$

which can be incorporated into the solute Hamiltonian to model the electrostatic, polarising influence of the solvent on the solute. In practice, the 3D-RISM equations are solved on a grid and it is therefore practical to also discretise the solvent charge density on the same grid by mapping it to voxel-centred point charges

$$q_{\text{solv},i}(\mathbf{r}_i) = \rho_q(\mathbf{r}_i)\Delta V, \tag{4.48}$$

where $\Delta V = \Delta x \Delta y \Delta z$ is the volume of the voxel and $\mathbf{r}_i$ are the spatial coordinates of the voxel centres. Although formally a distribution function is obtained for each solvent site, the resulting solvent charges are obtained by superposition of all sites.[3] In the following a single solvent charge is denoted as $q_{\text{solv},i}$ as in the last equation and the complete set of charges is denoted as

$$q_{\text{solv}} = \{q_{\text{solv},i}(\mathbf{r}_i) | i \in 1, \ldots, N_{\text{solv}}\}, \tag{4.49}$$

where N_{solv} is the total number of embedding charges. Note that these charges implicitly depend on φ_{sol}, which will become more important later, when the ONIOM-EC-RISM model is derived.

The solvent charges can now be incorporated into the solute Hamiltonian. In the context of EC-RISM it is advantageous to split the total Hamiltonian

$$\hat{H}_{\text{tot}} = \hat{H}_1 + \hat{H}_2 \tag{4.50}$$

of the system into a term $\hat{H}_1$ that includes all intra-solute interactions and the remaining solvent charge interactions $\hat{H}_2$. The quantum mechanical energy of the solute from eq. 4.45 can thus be identified as the energy associated with the evaluation of the Hamiltonian on the total wave function of the system

$$E_{\text{sol}}(q_{\text{solv}}) = \langle \Psi_{\text{tot}} | \hat{H}_1 | \Psi_{\text{tot}} \rangle . \tag{4.51}$$

Similarly, the solute energy can be split into the terms

$$E_{\text{sol}}(q_{\text{solv}}) = E_{\text{tot}}(q_{\text{solv}}) - E_2(q_{\text{solv}}), \tag{4.52}$$

where the energy E_1 has been renamed E_{sol} and

$$E_2(q_{\text{solv}}) = E_{\text{self}}(q_{\text{solv}}) + E_q(q_{\text{solv}}) \tag{4.53}$$

can be further split into the self-interaction energy of the charges E_{self} and the solvent-solute interaction energy E_q.[3] Depending on the technical implementation, a route via the direct evaluation of the $\hat{H}_1$-bra-ket or the explicit calculation of E_{tot} and E_2 may be available.

Similarly, the total free energy is defined as

$$G_{\mathrm{tot}} = G_{\mathrm{sol}} + E_2(q_{\mathrm{solv}}) = E_{\mathrm{sol}}(q_{\mathrm{solv}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}, u_{\mathrm{LJ}}) + E_2(q_{\mathrm{solv}}). \tag{4.54}$$

Formally, the free energies G_{tot} and G_{sol} are also functions of q_{solv}, φ_{sol} and u_{LJ}, but the functional arguments for these quantities are omitted for ease of notation.

As briefly mentioned above, the calculation of the excess chemical potential and the solvent charges depends on the solute potential, which in turn is polarised by the solute charges, giving rise to equations that must be solved iteratively. The technical implementation of the EC-RISM solver is discussed in section 4.5.3.

4.5.2. A note on EC-RISM, 3D-RISM-SCF and related methods

EC-RISM is a variant of a family of methods that can be broadly described as RISM-SCF, the most prominent of which is 3D-RISM-SCF. However, in the literature EC-RISM has also been referred to as 3D-RISM-SCF,[101] which may lead to confusion. For clarification, a brief timeline of the development of these adjecent methods will be given.

RISM-SCF models combine a RISM-based description of the solvent with a quantum mechanical description of the solute. The first model in this family was developed by Ten-no et. al[102] and was later extended to allow the calculation of free energies.[103] This method, called "1D-RISM self-consistent field" (1D-RISM-SCF), was also extended to the three-dimensional case (3D-RISM-SCF)[104] after the development of 3D-RISM. This approach requires direct integration of the RISM procedure into the QM code and was initially limited to Hartree-Fock calculations. In 2003, Du et al. presented an embedded cluster formulation of this method, called QM/3D-RISM-HNC, which first uses a grid of solvent charges in which the solute is embedded.[105,106] Later, independently, Kloss et al. used a similar approach to formulate the EC-RISM method. The authors note that Du et al. had not yet developed their QM/3D-RISM-HNC model into a free energy prediction tool.[3] EC-RISM must therefore be seen as a further development in this respect. These embedded cluster-type RISM models have the advantage that they can be implemented in principle with any QM code that allows the integration of the solute charges into the Hamiltonian, and thus do not require direct integration into the QM code, and therefore offer greater flexibility in terms of the levels of theory available. As a result, EC-RISM does not rely on a variational formalism for the electronic and free energies, and in practice EC-RISM is equivalent to 3D-RISM-SCF for pure variational methods.

Recently, Sato and coworkers have published a paper[107] that may provide the interested reader with a more detailed overview of these adjecent models.

4.5.3. Technical implementation of the EC-RISM method

The interdependence created by the mutual solute-solvent polarisation requires an iterative solution. The general EC-RISM procedure for calculating the solute free energy can be divided into three steps: An initial guess for the solvent distribution, followed by the main EC-RISM iterations where the free energy of the solute is converged. The last step is a final iteration that gives the final estimate of G_{sol}. These steps are visually summarised in figure 4.1 and will be further explained below.

The initial guess

The EC-RISM method consists of two basic building blocks, the first being the quantum mechanical calculation of the solute in the set of solvent charges, from which E_{sol} and φ_{sol} are obtained. The second building block is the calculation of the solvent structure, which takes the solute ESP and returns the set of solvent charges as well as the excess chemical potential. As both parts depend on the output of the other, it is obvious that the iterative EC-RISM procedure must be initialised with a reasonable first guess, either of the solute ESP or of the solvent charges. The first option can be achieved by omitting the solvent charges and thus performing a vacuum calculation to obtain a first estimate of the solute ESP. Alternatively, a first guess can replace the initial vacuum calculation with a PCM calculation that gives the ESP of an already solvated molecule. This procedure may be feasible if the vacuum calculation is difficult to converge and may also be considered as a first estimate of the solvent charges via a different but conceptually similar solvation model. Most commonly, a vacuum first guess is chosen.

In both cases the resulting ESP is used as input to the 3D-RISM solver. This general sequence of QM calculation of the solute and subsequent calculation of the solvent structure is referred to as an EC-RISM iteration and can be seen as the EC-RISM analogue of the iterations used in PCM double iteration scheme described in section 3.3.1.

The main EC-RISM iterations

After the initial guess, the solute free energy is converged by EC-RISM iterations. The set of solvent charges from one EC-RISM iteration is used to initialise the next iteration until it satisfies the convergence criteria

$$|G_{\mathrm{tot},k} - G_{\mathrm{tot},k-1}| < \epsilon_{\mathrm{tol}} \tag{4.55}$$

for a sufficently small ϵ_{tol} in iteration k. Here, G_{tot} is used to estimate the convergence, as the evaluation of E_2 is very expensive, especially for large lattices, and is usually restricted to the final iteration. As E_{tot} includes not only the solute-solvent interaction E_q but also the self-interaction energy E_{self}, the costly evaluation of the latter is often suppressed in electronic structure codes.

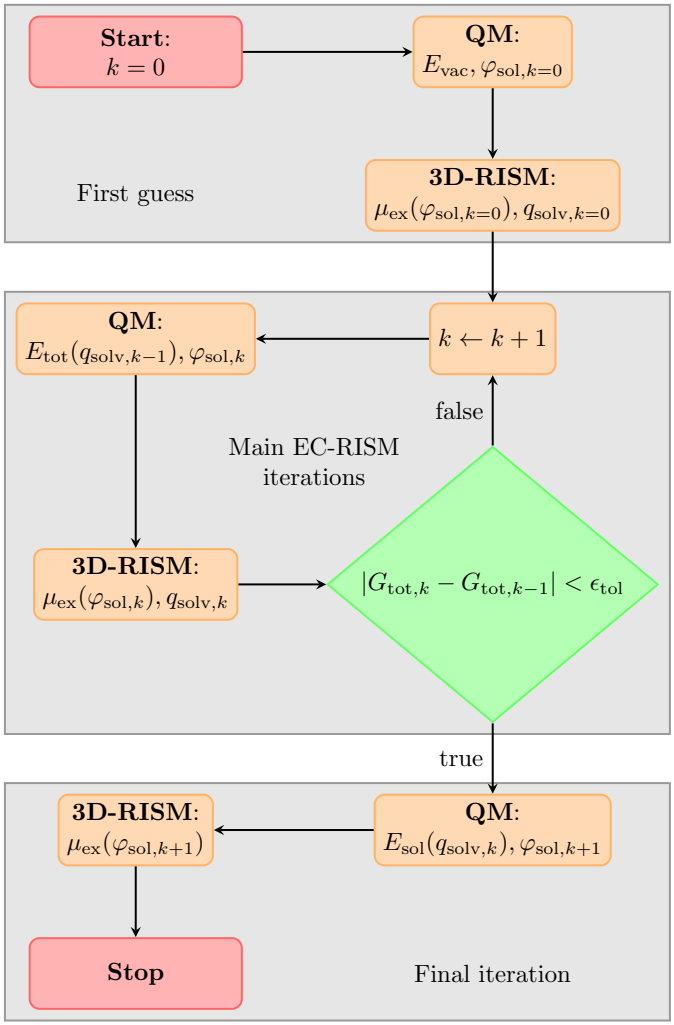

Figure 4.1.: Flowchart of a standard EC-RISM calculation. It is assumed that the QM code does not provide the solute-solvent-charge interaction E_2, so G_{tot} is used to estimate the convergence of the main EC-RISM iterations. The expensive calculation of E_2 from the ESP is only performed in the final iteration, which provides the final estimate for G_{sol}. The quantities returned by each process are given below the heading in each node, while the function arguments indicate the required inputs. Here a vacuum first guess is used as this is the standard procedure within EC-RISM, but an initialisation of the iterations using a PCM solvent is also available.

In order to further reduce the overall computational cost associated with reading the solvent charges through the QM code, Jochen Heil developed a charge compression scheme based on a Voronoi decomposition, which allows the number of point charges in the QM calculation to be drastically reduced.[17] This charge compression scheme is used throughout this thesis and is executed at the beginning of each EC-RISM cycle, except for the initial guess.

The main iterations account for the majority of the computational time required for EC-RISM to converge. It is therefore desirable to further reduce the cost of this part of the iterative process. For the commonly used MP2 level of theory and other post-Hartree-Fock theories, this is done by performing Hartree-Fock calculations in the main EC-RISM iterations and evaluating the perturbation terms only in the final iteration.

In practice, the EC-RISM procedure has been implemented using a number of QM codes, the most important for this work being EMPIRE[108] and ORCA.[18,19] EMPIRE, which is designed for SQM calculations, allows the term E_q to be read directly from the output, while the calculation of E_{self} is completely suppressed for performance reasons. In this case it is possible to use G_{sol} instead of G_{tot} to estimate convergence without any loss of performance.

In the case of the EC-RISM implementation with ORCA, the calculation of the self-interaction energy of the solvent charges is also suppressed, but E_q has to be calculated from the solvent charges and the solute ESP. The costly calculation of the latter in each main iteration is avoided, so G_{tot} is used to estimate convergence.

The final EC-RISM iteration

This part of the EC-RISM procedure generally includes all the calculations required for the final estimates of the thermodynamic quantities, in particular G_{sol}, by evaluating the necessary terms that were omitted in the main iterations for performance reasons. This includes the calculation of E_2 for codes that do not provide it directly in the output. This part of the EC-RISM procedure will become more important in the ONIOM-EC-RISM calculations. In addition, spectroscopic parameters can be calculated from the fully solvated wave function.

4.5.4. The PMV-correction: Empirical corrections to the free energy

Although the 3D-RISM based models provide an advanced description of the solvent structure, it has been shown that the application of corrections is necessary to obtain accurate thermodynamic quantities, in particular solvation free energies (SFEs),[23,24,109–112] which are commonly used to estimate the performance of solvation models. Part of the error can be attributed to the overestimation of the pressure required to form the solvent cavity, which can be corrected by partial molar volume terms.[100]

Combining the correction to the PMV with a second term for solutes with the charge

q_{sol}, which attempts to correct for the neglect of the Galvani potential in the determination of the hydration free energy of the proton,[23,113] Tielker et al. formulated a linear correction

$$G_{\mathrm{sol}}^{\mathrm{corr}} = E_{\mathrm{sol}}(q_{\mathrm{solv}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}, u_{\mathrm{LJ}}) + c_V V_{\mathrm{m}}(\varphi_{\mathrm{sol}}, u_{\mathrm{LJ}}) + c_q q_{\mathrm{sol}} \tag{4.56}$$

to the solute free energy obtained from EC-RISM.[23] For simplicity, this two-parameter model is referred to as the PMV correction. The parameters c_V and c_q are obtained by fitting the calculated SFEs to experimental values. It should be noted that as a consequence, the empirically corrected free energies implicitly show a dependence on the fitted vacuum energies, regardless of whether the physical quantity derived from them explicitly includes vacuum energy terms. Note that equation 4.56 corrects the energy of a single solute conformer. See chapter 5 for details of the parameterisation approach and its application to ensembles of conformers.

4.6. ONIOM-EC-RISM

It is now finally time to outline the ONIOM-EC-RISM model. This task consists of two parts: The approximation of the quantum mechanical energy of the solute and the approximation of its ESP. For the derivation and subsequent references, it will be useful to reformulate equations 4.45 and 4.56 in ONIOM notation as

$$G_{\mathrm{sol}}^{\mathrm{hr}} = E_{\mathrm{sol}}^{\mathrm{hr}}(q_{\mathrm{solv}}^{\mathrm{hr}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}^{\mathrm{hr}}, u_{\mathrm{LJ}}), \tag{4.57}$$

and

$$G_{\mathrm{sol}}^{\mathrm{hr,corr}} = G_{\mathrm{sol}}^{\mathrm{hr}} + c_V^{\mathrm{hr}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{hr}}, u_{\mathrm{LJ}}) + c_q^{\mathrm{hr}} q_{\mathrm{sol}} \tag{4.58}$$

which will be the respective *hr*-references and approximation targets for the basic ONIOM-EC-RISM schemes and the PMV-corrected models. The approximations used here are similar to those used in the ONIOM-PCM schemes. Therefore, we will start with the EC-RISM analogue of the /A scheme.

4.6.1. ONIOM-EC-RISM/A

In the ONIOM-EC-RISM/A scheme, the solute energy is approximated by an ONIOM calculation performed in a single set of solvent charges. In order to calculate the solvent structure and subsequently the solvent charges, the electrostatic and Lennard-Jones contributions to the solute-solvent interaction potential must be transferred into the ONIOM context.

Using the basic assumption of the ONIOM-PCM/A model of Vreven et al.[11] that the solute charge density can be expressed as the ONIOM extrapolation

$$\rho_{\mathrm{sol}}^{\mathrm{ONIOM}} = \Omega(\boldsymbol{\rho}_{\mathrm{sol}}) = \rho_{\mathrm{sol}}^{\mathrm{lr}} + \rho_{\mathrm{sol}}^{\mathrm{hm}} - \rho_{\mathrm{sol}}^{\mathrm{lm}} =: \rho_{\mathrm{sol}}^{/\mathrm{A}}, \tag{4.59}$$

the ONIOM-ESP can also be expressed as

$$\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}} = \frac{1}{4\pi\varepsilon_0} \int \frac{\Omega(\boldsymbol{\rho}_{\mathrm{sol}})}{\|\mathbf{r} - \mathbf{r}'\|} \, \mathrm{d}\mathbf{r}' = \Omega(\boldsymbol{\varphi}_{\mathrm{sol}}) = \varphi_{\mathrm{sol}}^{\mathrm{lr}} + \varphi_{\mathrm{sol}}^{\mathrm{hm}} - \varphi_{\mathrm{sol}}^{\mathrm{lm}}. \tag{4.60}$$

Here the additivity of the integral and the ONIOM extrapolation has been applied.

The second part of the solute-solvent interaction potential $u_{\mathrm{LJ}}^{\mathrm{ONIOM}}$ is derived using a standard force field approach, i.e. the chosen LJ parameters depend only on the molecular connectivity. It is therefore true that

$$u_{\mathrm{LJ}}^{\mathrm{hm}} = u_{\mathrm{LJ}}^{\mathrm{lm}} \tag{4.61}$$

and consequently

$$u_{\mathrm{LJ}} \coloneqq u_{\mathrm{LJ}}^{\mathrm{ONIOM}} = \Omega(\mathbf{u}_{\mathrm{LJ}}) = u_{\mathrm{LJ}}^{\mathrm{lr}} = u_{\mathrm{LJ}}^{\mathrm{hr}}. \tag{4.62}$$

This means that the ONIOM approximation of the LJ potential is exact and will be dropped as a function argument when referring to the solute-solvent interaction potential, as it is not required for the further discussion of the ONIOM-EC-RISM method. The set of solvent charges are therefore shortly written

$$q_{\mathrm{solv}}^{\mathrm{ONIOM}} \coloneqq q_{\mathrm{solv}}(\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}}), \tag{4.63}$$

where the solute potential can also be omitted for brevity. This set of solvent charges is derived in an analogous manner to the standard *hr*-EC-RISM procedure (equation 4.49), as the ONIOM-ESP $\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}}$ and u_{LJ} are now used as input to the 3D-RISM solver instead of the *hr*-ESP $\varphi_{\mathrm{sol}}^{\mathrm{hr}}$, resulting in a solvent distribution that is dependent on this modified interaction potential.

In addition to the solvent charges, the *model* system can be polarised in both *hm* and *lm* calculations by the residual solute environment if an electronic embedding scheme is employed. The complete set of polarising charges is therefore defined as

$$q^{\mathrm{ONIOM}} = q_{\mathrm{solv}}^{\mathrm{ONIOM}} \cup q_{\mathrm{EE}}^{\mathrm{lr}} \tag{4.64}$$

where $q_{\mathrm{EE}}^{\mathrm{lr}} = \{\}$ in the case of mechanical embedding.

Using these charges, the ONIOM approximation of the solute energy is defined as

$$\begin{aligned} E_{\mathrm{sol}}^{/\mathrm{A}}(q^{\mathrm{ONIOM}}) &= \Omega(\mathbf{E}_{\mathrm{sol}}(q^{\mathrm{ONIOM}})) \\ &= E_{\mathrm{sol}}^{\mathrm{lr}}(q^{\mathrm{ONIOM}}) + E_{\mathrm{sol}}^{\mathrm{hm}}(q^{\mathrm{ONIOM}}) - E_{\mathrm{sol}}^{\mathrm{lm}}(q^{\mathrm{ONIOM}}) \end{aligned} \tag{4.65}$$

in analogy to the same quantity in the *hr*-reference from equation 4.57. As outlined in the previous sections for the basic EC-RISM model, this quantity may only be available via the route of the total energy of the system and the explicit calculation of the interactions

associated with the solvent charges. In the context of the ONIOM-EC-RISM model, equation 4.52 must therefore be reformulated as

$$E_{\mathrm{sol}}^{/\mathrm{A}}(q^{\mathrm{ONIOM}}) = \Omega(\mathbf{E}_{\mathrm{tot}} - \mathbf{E}_2) = \Omega(\mathbf{E}_{\mathrm{tot}} - (\mathbf{E}_q + \mathbf{E}_{\mathrm{self}})). \tag{4.66}$$

Here the explicit dependence of $\Omega(\mathbf{E}_{\mathrm{tot}})$, $\Omega(\mathbf{E}_2)$ and its contributing terms on q^{ONIOM} is omitted to shorten the equation.

To reduce the overall cost, the calculation of the self-interaction energy of the solvent charges is suppressed in most codes, while E_q^{ONIOM} can be calculated explicitly with

$$E_q^{\mathrm{ONIOM}} = \Omega(\mathbf{E}_q) = \sum_{q_i \in q_{\mathrm{solv}}^{\mathrm{ONIOM}}} q_i \varphi_i^{\mathrm{lr}}(\mathbf{r}_i) + \sum_{q_j \in q^{\mathrm{ONIOM}}} q_j \left(\varphi_j^{\mathrm{hm}}(\mathbf{r}_j) - \varphi_j^{\mathrm{lm}}(\mathbf{r}_j) \right). \tag{4.67}$$

The electrostatic solute potential is evaluated at the spatial coordinates of the solvent charges. This equation is also valid for all other ONIOM-EC-RISM schemes presented in this chapter, except for /X, where three independent EC-RISM calculations are performed and therefore the standard way of calculating E_2 as in *hr* is applied.

Using the ONIOM-ESP for calculating the excess chemical potential and the set of point charges, and evaluating the electronic energy of the solute within this point charge field, we obtain the free energy of the solute

$$G_{\mathrm{sol}}^{/\mathrm{A}} = E_{\mathrm{sol}}^{/\mathrm{A}}(q^{\mathrm{ONIOM}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}}), \tag{4.68}$$

within the ONIOM-EC-RISM/A framework. As with the *hr*-EC-RISM method, this equation requires an iterative solution, the technical implementation of which is described below.

Technical implementation

The ONIOM-EC-RISM/A model presented so far is general with respect to the chosen theory levels *low* and *real* as well as the codes used to perform the calculation of E_{sol} or E_{tot}. In order to discuss the specifics of the implementation, this generalisation is dropped.

The /A scheme and all other ONIOM-EC-RISM schemes presented later are implemented using the QM software ORCA 5[18,19] for the *high* level calculations and EMPIRE20[108,114] for the *low* level calculations. These codes were chosen on the basis of availability and performance. The available combinations of methods are therefore limited to those implemented in ORCA and EMPIRE. The former offers a wide range of QM methods, while the latter offers SQM methods, so the following discussion will be concerned with ONIOM2(QM:SQM) calculations. In principle, it would have been possible to use ORCA for all sub-calculations, as SQM methods are also available in this code, but they are currently not parallelised. EMPIRE was chosen for the *lr* and

lm computations to allow larger systems to be calculated in a reasonable time, as it is highly parallelised.[114] A parallelised in-house code was used for the 3D-RISM calculations. The EE-charges $q^{\mathrm{lr}}_{\mathrm{EE}}$ are read as Coulson charges from the EMPIRE output. The generalised flowchart of the ONIOM-EC-RISM/A implementation with ORCA and EMPIRE is shown in figure 4.2.

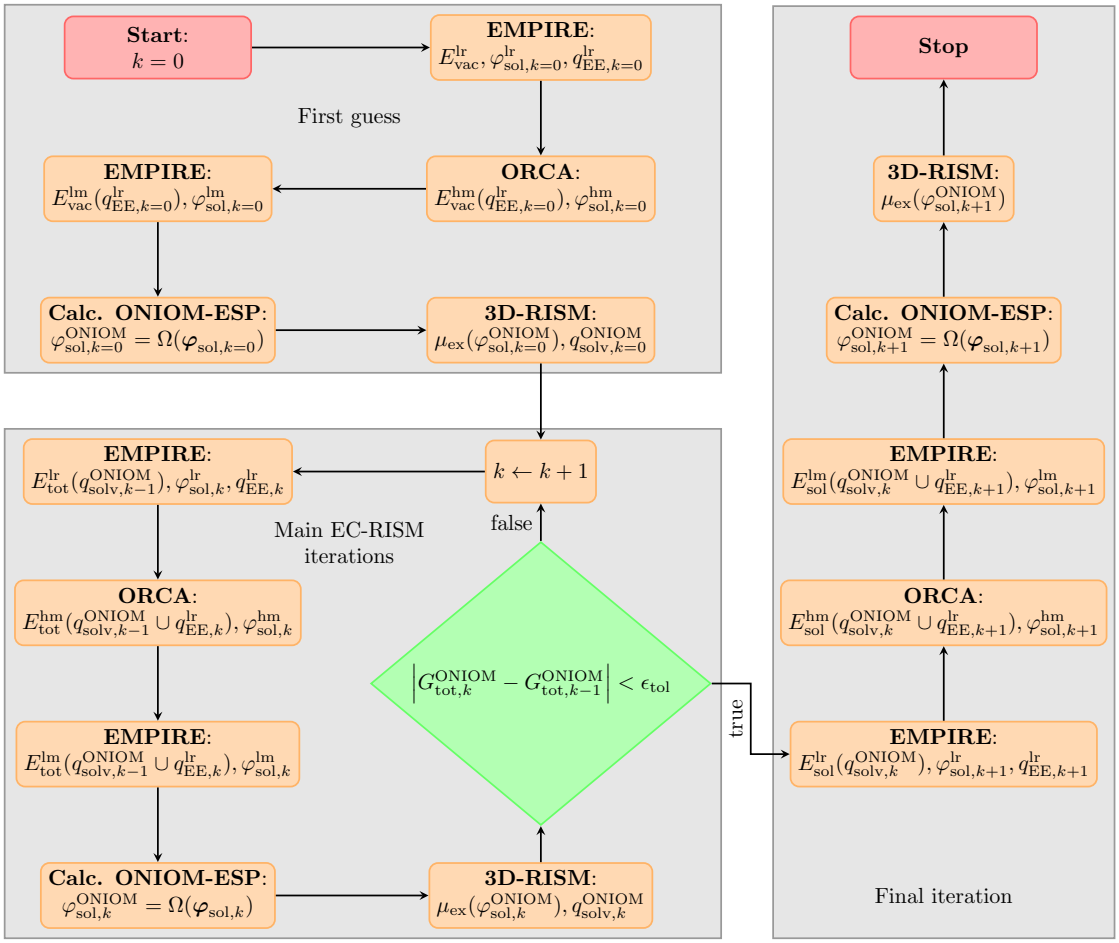

Figure 4.2.: Flowchart of the ONIOM-EC-RISM/A implementation. The presented notation is general with respect to ME and EE, since $q^{\mathrm{ONIOM}} = q^{\mathrm{ONIOM}}_{\mathrm{solv}}$ for ME.

In general, the technical implementation is similar to *hr*-EC-RISM. The process starts with an ONIOM calculation in vacuum. This provides a first estimate of the solute ONIOM-ESP. Alternatively, the initial *lr* vacuum calculation can be replaced by a PCM calculation, which allows to accelerate the convergence of the *lr* calculation for larger protein systems with charged surfaces. The difficulty of converging these types of systems in vacuum was shown by Wick et al.[115] The ONIOM-ESP is then fed to the

3D-RISM solver and the resulting solvent charges are used to initialise the main EC-RISM iterations.

After $G_{\mathrm{tot}}^{/\mathrm{A}}$ has been converged in the main iterations, the final iteration is carried out with the explicit calculation of $E_{\mathrm{sol}}^{/\mathrm{A}}$. EMPIRE provides the solute-point-charge interaction in the log file, while the interaction with the *hm* charge density from ORCA is calculated explicitly from the *hm*-ESP, using an expression similar to equation 4.67. The calculation of E_{self} is suppressed in EMPIRE and avoided in ORCA by specifying the solvent and EE charges in a point charge file separate from the input file.

In this implementation, the *model* system is constructed by specifying the subset of *model* system atoms by their indices. The bond atoms are placed using the method of bond scaling presented by Vreven et al. (equation 2.8). The required scaling parameters are taken from the ONIOM implementation in Gaussian16. They are available in the electronic supplementary material.[116] This scaling method is used for all ONIOM-EC-RISM schemes presented in this work.

All calculations with EMPIRE are performed with the "Randomize" keyword set to zero, turning off randomisation of the initial guess matrix, as it was observed that this would otherwise lead to divergent free energy predictions for repeated calculations of the same molecule with identical EC-RISM settings. Converged wave functions from the previous EC-RISM iteration are read as a first guess for the new wave function to speed up convergence. In addition, an analogue scheme has been implemented for 3D-RISM calculations where the *c*-function from the previous calculation is read by the 3D-RISM solver to initialise the current solvent structure calculation.

PMV-correction and extrapolation limit

Since it has been demonstrated that the *hr*-EC-RISM energies require a correction based on both the solute's partial molar volume and charge,[23,24] it can be anticipated that a correction to the free energies of the ONIOM-EC-RISM approximations will also be necessary. Therefore, the PMV correction presented in section 4.5.4 must be transferred to the ONIOM-EC-RISM framework.

In the case of the /A model this is straightforward as the expression for the solute free energy is similar to that of the *hr*-approach. Using the PMV obtained from the ONIOM-ESP, the correction is defined as

$$G_{\mathrm{sol}}^{/\mathrm{A,corr}} = G_{\mathrm{sol}}^{/\mathrm{A}} + c_V^{/\mathrm{A}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}}) + c_q^{/\mathrm{A}} q_{\mathrm{sol}}, \tag{4.69}$$

where the charge of the solute $q_{\mathrm{sol}} \coloneqq \Omega(\mathbf{q}_{\mathrm{sol}}) = q_{\mathrm{sol}}^{\mathrm{lr}} = q_{\mathrm{sol}}^{\mathrm{hr}}$ is used for the charge correction term.

A consistent parameterisation of this correction would require the ONIOM method to be applied to a training set of partitioned molecules. However, the size of these datasets prevents manual partitioning of their molecules in most cases. In addition, a procedure

would have to be found that minimises the partitioning error, since manual partitioning would introduce an additional arbitrary error.

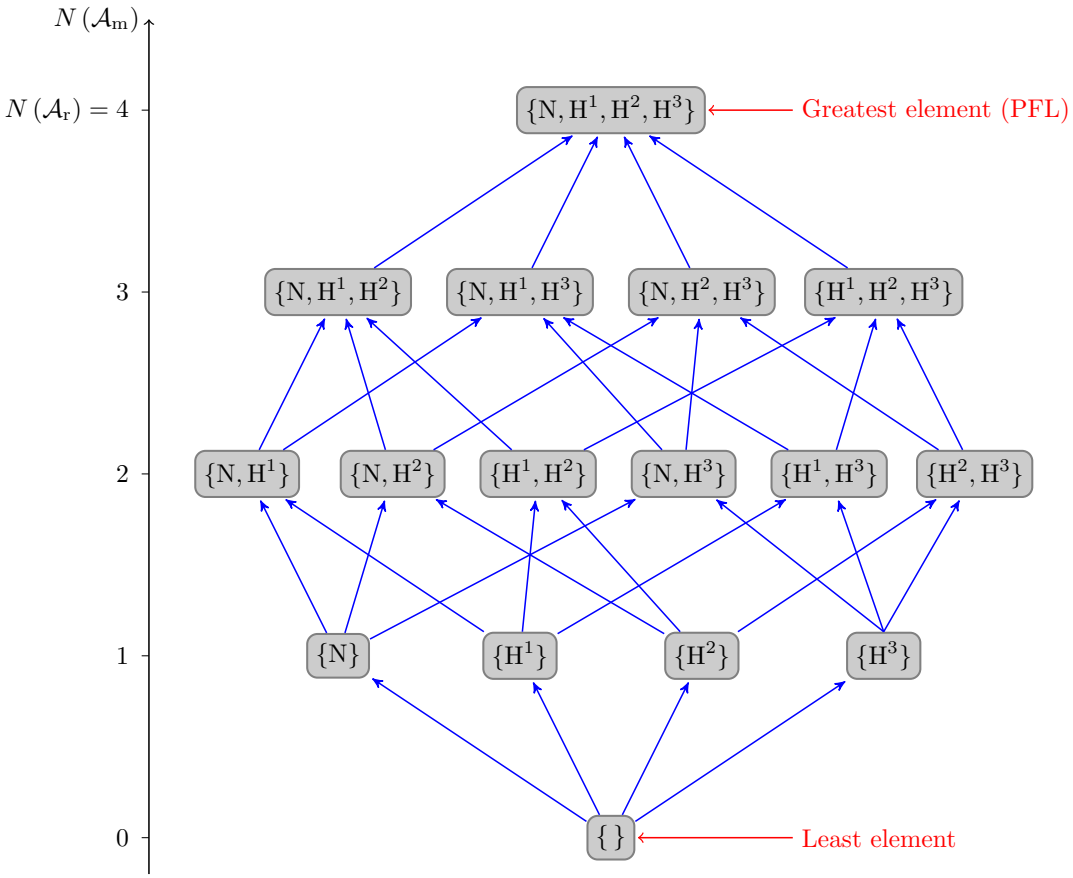

Figure 4.3.: Hasse diagram of all possible *model* systems $\mathcal{A}_\mathrm{m}$, i.e. the power set of $\mathcal{A}_\mathrm{r}$, ordered by inclusion for an ammonia molecule. The axis shows the number of atoms in the *model* system. For brevity, the set of link atoms is excluded from $\mathcal{A}_\mathrm{m}$. The greatest element can be used to parameterise a PMV correction that does not require partitioning of the molecules in the training dataset.

A solution to this problem lies in the application of partition-free corrections: When an ONIOM partition is applied to a molecular system, i.e. the *real* system, it is divided into two subsets, $\mathcal{A}_\mathrm{r}$ and $\mathcal{A}_\mathrm{m}$, as previously outlined in section 2.2. The latter is used to define the *model* system by adding link atoms. Suppose we partially order the set of all possible *model* system sets by inclusion or size, we then obtain a set with two extremes. The smallest element is the empty set, where there are no atoms in the *model* system. Conversely, the largest element represents the case where all atoms of the *real* system are

part of the *model* system and thus represents the size extrapolation limit of the ONIOM method. We will refer to this element as the Partition Free Limit (PFL), as there is no ONIOM boundary and therefore no partition when using this largest element. This process is illustrated in figure 4.3 for an ammonia molecule.

All quantities derived within this limit are free of any ONIOM partitioning error, which can be used for the PMV correction. Applying this element to the ONIOM-EC-RISM/A scheme, and consequently the associated PMV correction, gives the expression for *hr*-EC-RISM (equations 4.57 and 4.58).

It is expected that if the extrapolation to the *hr*-target, as defined by equation 4.69, works, the application of the *hr* correction parameters is justified and no additional PMV correction needs to be parameterised. In other words, c_V^{hr} and c_q^{hr} are sufficient approximations for $c_V^{/\mathrm{A}}$ and $c_q^{/\mathrm{A}}$. This hypothesis will be investigated in the following chapters.

4.6.2. ONIOM-EC-RISM/B

The ONIOM-EC-RISM/B scheme is based on the approximation of $\rho_{\mathrm{sol}}^{\mathrm{ONIOM}}$ by

$$\rho_{\mathrm{sol}}^{/\mathrm{B}} \coloneqq \rho_{\mathrm{sol}}^{\mathrm{lr}}. \tag{4.70}$$

Hence, from equation 4.60 one directly obtains the *lr*-ESP $\varphi_{\mathrm{sol}}^{\mathrm{lr}}$ as an approximation of $\varphi_{\mathrm{sol}}^{\mathrm{ONIOM}}$.

Inserting this ESP into equation 4.65 yields the solute electronic energy of the /B approximation

$$E_{\mathrm{sol}}^{/\mathrm{B}}(q^{\mathrm{lr}}) = \Omega(\mathbf{E}_{\mathrm{sol}}(q^{\mathrm{lr}})), \tag{4.71}$$

where the set of solvent charges

$$q_{\mathrm{solv}}^{\mathrm{lr}} \coloneqq q_{\mathrm{solv}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) \tag{4.72}$$

is derived using the *lr*-ESP $\varphi_{\mathrm{sol}}^{\mathrm{lr}}$ and can be extended by EE charges in the same way as in the /A scheme

$$q^{\mathrm{lr}} = q_{\mathrm{solv}}^{\mathrm{lr}} \cup q_{\mathrm{EE}}^{\mathrm{lr}}. \tag{4.73}$$

The interaction of the solute with the polarising charges can be calculated using an expression analogous to equation 4.67 with this set of charges.

The excess chemical potential can be approximated in the same way using the *lr*-ESP, which gives the expression for the free energy of the solute

$$G_{\mathrm{sol}}^{/\mathrm{B}} = E_{\mathrm{sol}}^{/\mathrm{B}}(q^{\mathrm{lr}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}). \tag{4.74}$$

This equation can be rewritten as

$$G_{\mathrm{sol}}^{/\mathrm{B}} = G_{\mathrm{sol}}^{\mathrm{lr}} + E_{\mathrm{sol}}^{\mathrm{hm}}(q^{\mathrm{lr}}) - E_{\mathrm{sol}}^{\mathrm{lm}}(q^{\mathrm{lr}}), \tag{4.75}$$

which shows the main advantage of the ONIOM-EC-RISM/B approximation: The solvent structure calculation can be completely decoupled from the *model* system calculations, as in the ONIOM-PCM/B scheme, reducing the number of evaluations of the expansive $E_{\mathrm{sol}}^{\mathrm{hm}}$ term to one. The free energy of the solute is thus the ONIOM extrapolation of $G_{\mathrm{sol}}^{\mathrm{lr}}$ from a standard EC-RISM calculation at the *low* level of theory and the *model* system energies evaluated once with the already converged solvent distribution.

Technical implementation

The implementation of the ONIOM-EC-RISM/B method is similar to that of the /A method and only the differences will be outlined below. The corresponding flow chart is shown in figure 4.4.

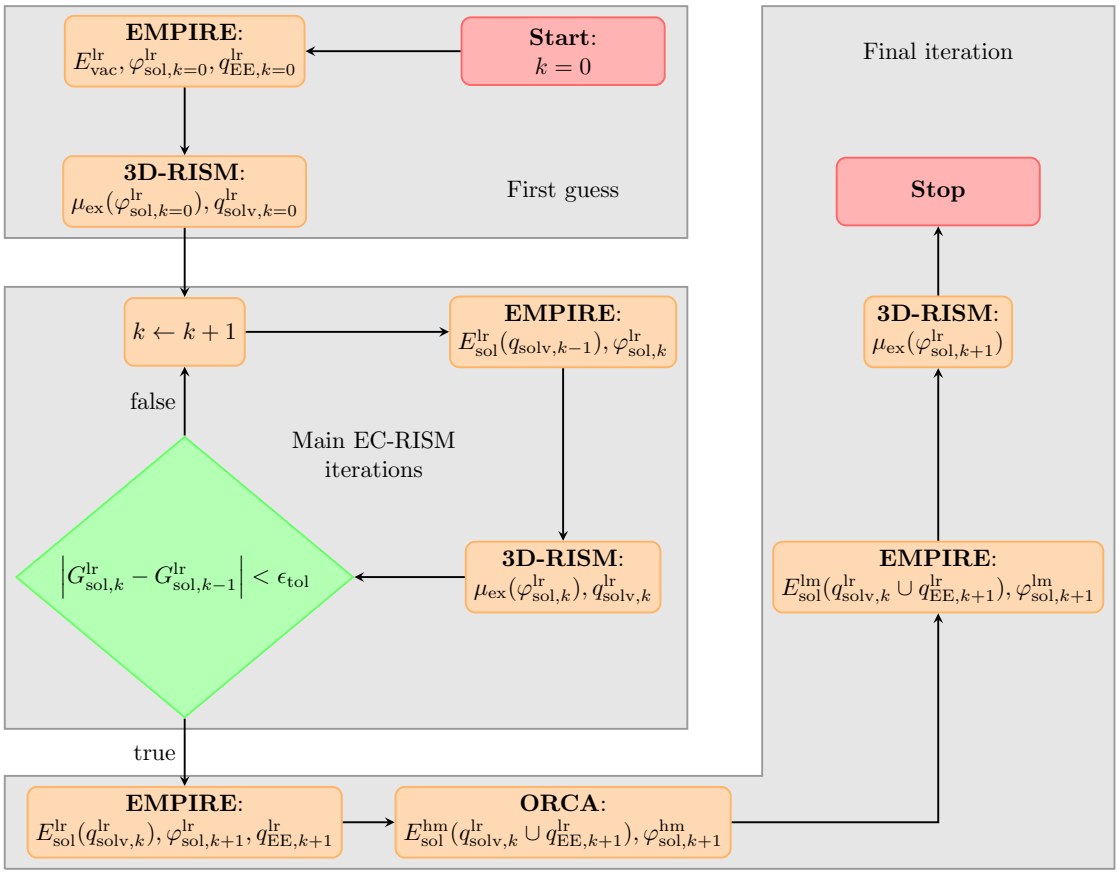

Figure 4.4.: Flowchart of the ONIOM-EC-RISM/B implementation.

As the solvent structure depends only on the ESP from the *lr* calculation and is therefore independent of the output from the *model* system calculations, $G_{\mathrm{sol}}^{\mathrm{lr}}$ can be

converged first. This is done in a similar way to the *hr*-EC-RISM scheme: First a first guess for the *real* system in vacuum or PCM solvent is made at the *low* level of theory. The main iterations are performed until $G_{\mathrm{sol}}^{\mathrm{lr}}$ converges.

In the final iteration, the converged solvent charge field can be used to perform the *model* system calculations. In contrast to the /A scheme, the extensive *hm* calculations only need to be performed once instead of once per iteration, which greatly reduces the overall computational cost.

This scheme is the EC-RISM analogue of the ONIOM-PCM double iteration scheme which motivated the development of the more approximate schemes. At present there is no equivalent implementation for the single iteration scheme of PCM, therefore the expected reduction in computational cost for the ONIOM-EC-RISM/B scheme is that inherent in the approximation of the ONIOM-ESP by the *lr*-ESP. However, it should be noted that efforts are currently being made to implement a single iteration scheme for *hr*-EC-RISM in the ORCA software package. This scheme can then be easily adopted for the ONIOM-EC-RISM/B scheme.

PMV-correction and extrapolation limit

In the same way as the PMV correction is formulated for the *hr* and /A scheme, it can be written as

$$G_{\mathrm{sol}}^{/\mathrm{B,corr}} = G_{\mathrm{sol}}^{/\mathrm{B}} + c_V^{/\mathrm{B}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) + c_q^{/\mathrm{B}} q_{\mathrm{sol}} \tag{4.76}$$

for the ONIOM-EC-RISM/B scheme, where V_{m} is calculated from the *lr*-ESP.

In contrast, applying the PFL to this scheme results in a size extrapolation limit

$$G_{\mathrm{sol}}^{/\mathrm{B@PFL}} = E_{\mathrm{sol}}^{\mathrm{hr}}(q^{\mathrm{lr}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) \tag{4.77}$$

that is different from the previous scheme. This is a mixed model where the *real* system is evaluated with the *high* level theory within a solvent charge field calculated from the *lr*-ESP. The solute is thus described at the *high* level, while the solvent is modelled at the cheaper *low* level, marking a halfway point between the *hr*-EC-RISM approach, where both are described at the *high* level, and the corresponding *lr*-calculation, where both are modelled at the *low* level of theory.

The PMV correction within this extrapolation limit is therefore

$$G_{\mathrm{sol}}^{/\mathrm{B@PFL,corr}} = E_{\mathrm{sol}}^{\mathrm{hr}}(q^{\mathrm{lr}}) + \mu_{\mathrm{ex}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) + c_V^{/\mathrm{B@PFL}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) + c_q^{/\mathrm{B@PFL}} q_{\mathrm{sol}}. \tag{4.78}$$

As with the other schemes the usage of PFL aims to give an approximation to the parameters that would be obtained from calculations on partitioned molecules. However, in contrast to the previously discussed /A scheme, the application of the PFL to /B yields a set of parameters that are not equal to those of the *hr*-PMV correction.

Nevertheless, it should be noted that if the *lr* method is able to accurately reproduce the *hr*-ESP, the /B@PFL scheme is equivalent to the *hr*-EC-RISM method and therefore

extrapolates to the same quantity as the /A scheme. If this is true, then the PMV correction at this limit is also equal to the corresponding *hr* correction. Therefore, both the *hr* and /B@PFL parameters may be an adequate approximation for the /B parameters, which will be tested in later parts of this work.

4.6.3. ONIOM-EC-RISM/C

For the sake of completeness, the ONIOM-PCM/C scheme is also transferred to the ONIOM-EC-RISM framework. In addition to the approximation of the ONIOM-ESP by the *lr*-ESP, the central approximation of the /B scheme, the solvent effect on the *model* system is neglected. This is achieved by evaluating the *model* systems only with the EE charges or without any additional charges in the case of ME.

The solvent charges are thus identical to the /B scheme, while the energy of the solute is given by

$$E_{\mathrm{sol}}^{/\mathrm{C}}(q^{\mathrm{lr}}) = E_{\mathrm{sol}}^{\mathrm{lr}}(q^{\mathrm{lr}}) + E_{\mathrm{vac}}^{\mathrm{hm}}(q_{\mathrm{EE}}^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{lm}}(q_{\mathrm{EE}}^{\mathrm{lr}}). \tag{4.79}$$

The expression for the free energy in the ONIOM-EC-RISM/C scheme is therefore given by

$$G_{\mathrm{sol}}^{/\mathrm{C}} = G_{\mathrm{sol}}^{\mathrm{lr}} + E_{\mathrm{vac}}^{\mathrm{hm}}(q_{\mathrm{EE}}^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{lm}}(q_{\mathrm{EE}}^{\mathrm{lr}}). \tag{4.80}$$

To evaluate the energetic difference between the /B and /C schemes one can compute the difference between equations 4.75 and 4.80, which gives

$$G_{\mathrm{sol}}^{/\mathrm{B}} - G_{\mathrm{sol}}^{/\mathrm{C}} = E_{\mathrm{sol}}^{\mathrm{hm}}(q^{\mathrm{lr}}) - E_{\mathrm{sol}}^{\mathrm{lm}}(q^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{hm}}(q_{\mathrm{EE}}^{\mathrm{lr}}) + E_{\mathrm{vac}}^{\mathrm{lm}}(q_{\mathrm{EE}}^{\mathrm{lr}}). \tag{4.81}$$

The two therefore give identical results if

$$E_{\mathrm{sol}}^{\mathrm{hm}}(q^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{hm}}(q_{\mathrm{EE}}^{\mathrm{lr}}) = E_{\mathrm{sol}}^{\mathrm{lm}}(q^{\mathrm{lr}}) - E_{\mathrm{vac}}^{\mathrm{lm}}(q_{\mathrm{EE}}^{\mathrm{lr}}), \tag{4.82}$$

which means that the energetic effect of the polarising solvent charges needs to be identical for the *low* and *high* level of theory, assuming that both /B and /C are calculated with the same embedding scheme, i.e. ME or EE. Otherwise the effect of the EE charges also has to be considered in the energy differences.

Technical implementation

Since the *model* system calculations are completely decoupled from the quantities obtained from the *lr* calculation in the ME scheme, no special implementation is required. The free energy of the solute can be obtained simply by calculating $G_{\mathrm{sol}}^{\mathrm{lr}}$ with the standard EC-RISM implementation and performing the ONIOM extrapolation with the *model* system energies in vacuum.

In contrast, in the EE model the *model* system calculations are still dependent on the *lr*-EE charges and therefore need to be performed after convergence of the solvent

distribution. As a consequence, the polarisation due to the solvent is transferred to the *model* system calculations via the EE charges. This slightly contradicts the basic approximation of the /C scheme, that the *model* system calculations should be decoupled from the solvent polarisation. One way around this problem is to move the *model* system calculations to the first guess iteration. However, this implies that the initial *lr* calculation must not be performed with a PCM solvent, which can lead to convergence problems for protein systems with charged surfaces. Due to these problems and since the goal of the development of the ONIOM-EC-RISM schemes is to reduce the overall computational cost, and the expected cost reduction compared to the /B scheme is limited, no specific effort was made to implement the EE variant of the /C scheme.

To illustrate the general idea behind the ONIOM-EC-RISM/C scheme, figure 4.5 shows a flowchart that is general in terms of the ME scheme and an EE scheme that uses the fully solvated *lr*-EE charges in the final iteration.

PMV-correction and extrapolation limit

The PMV correction for this scheme

$$G_{\mathrm{sol}}^{/\mathrm{C,corr}} = G_{\mathrm{sol}}^{/\mathrm{C}} + c_V^{/\mathrm{C}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) + c_q^{/\mathrm{C}} q_{\mathrm{sol}} \tag{4.83}$$

is very similar to that of the /B scheme. In contrast, the application of the PFL does not cause the energies from the lower level calculations to cancel out, and the resulting expression

$$G_{\mathrm{sol}}^{/\mathrm{C@PFL}} = G_{\mathrm{sol}}^{\mathrm{lr}} + E_{\mathrm{vac}}^{\mathrm{hr}} - E_{\mathrm{vac}}^{\mathrm{lr}} \tag{4.84}$$

contains the energies resulting from the evaluation of the *real* system at both levels of theory.

Similarly, the PMV correction for /C@PFL is

$$G_{\mathrm{sol}}^{/\mathrm{C@PFL,corr}} = G_{\mathrm{sol}}^{\mathrm{lr}} + E_{\mathrm{vac}}^{\mathrm{hr}} - E_{\mathrm{vac}}^{\mathrm{lr}} + c_V^{/\mathrm{C@PFL}} V_{\mathrm{m}}(\varphi_{\mathrm{sol}}^{\mathrm{lr}}) + c_q^{/\mathrm{C@PFL}} q_{\mathrm{sol}}. \tag{4.85}$$

4.6.4. ONIOM-EC-RISM/X

The last scheme is the simplest in theory and implementation. As in the ONIOM-PCM/X scheme, the free energy is approximated by using three sets of solvent charges. In the case of PCM, these are calculated using the individual solute charge densities from the sub-calculations, but are placed on the same cavity surface, as explained in detail in section 3.3.4.

In contrast to the PCM method, however, EC-RISM does not require the explicit specification of a solute cavity. Therefore, in the case of EC-RISM, the sub-calculations of the /X scheme are completely independent.

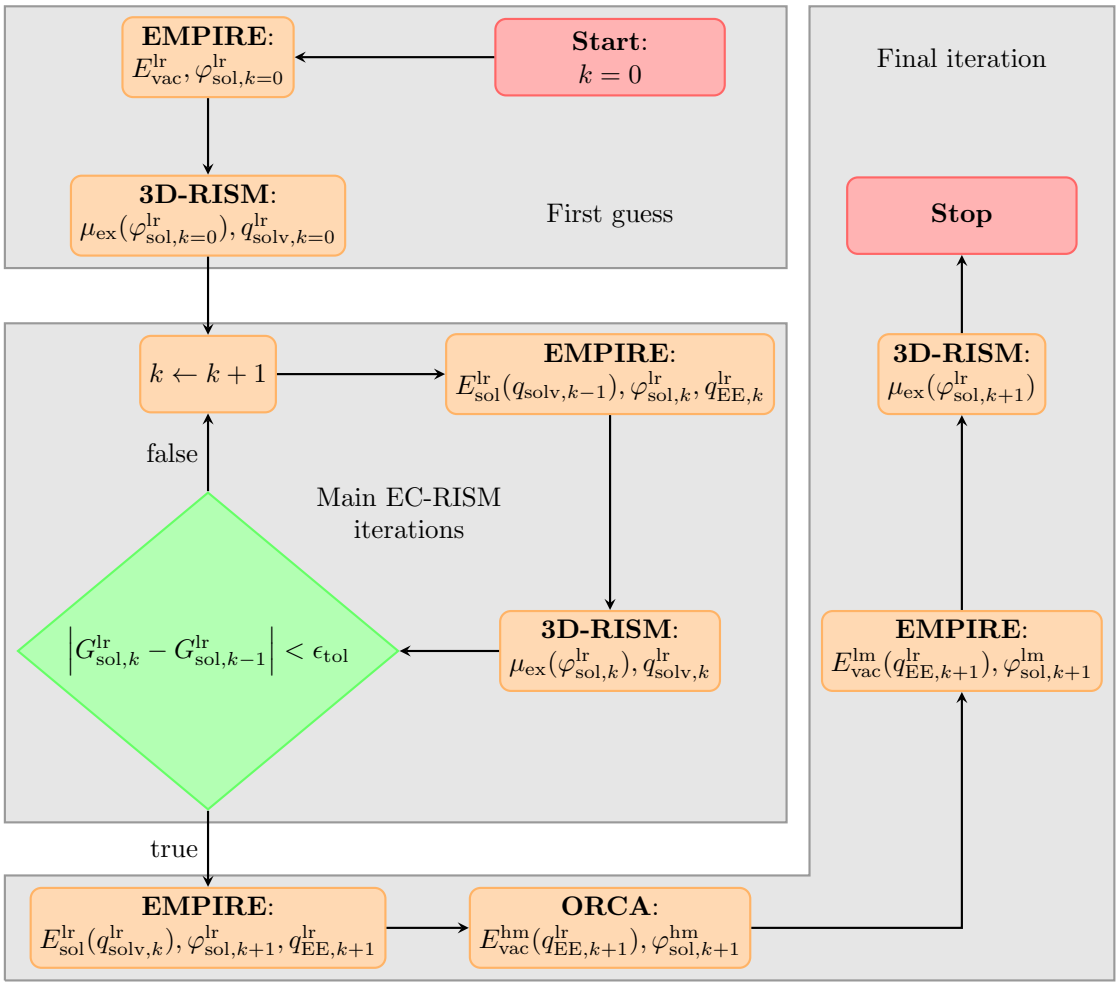

Figure 4.5.: Flowchart of the ONIOM-EC-RISM/C implementation. See the main text for the considerations that need to be made regarding ME and EE.

The free energy of the /X scheme is consequently given by

$$G_{\mathrm{sol}}^{/\mathrm{X}} = \Omega(\mathbf{G}_{\mathrm{sol}}), \tag{4.86}$$

where the elements of $\mathbf{G}_{\mathrm{sol}}$ are evaluated by independent EC-RISM calculations. This independence also implies that there are no embedding charges in the *model* system calculations. By construction, the ONIOM-EC-RISM/X scheme is therefore an ME scheme.

Technical implementation

Implementation of the scheme is straightforward. The coordinates of the *model* system are constructed from the *real* system with the given partitioning, and three calculations are performed using the standard EC-RISM implementation. The final estimate of the solute free energy is then obtained by extrapolation through the ONIOM extrapolation given in equation 4.86. In practice, the sub-calculations in this work are carried out using the same codes as for the other schemes, i.e. EMPIRE for the *low* level calculations and ORCA for the *hm* calculation, but since no additional modification is required for this scheme, it can be run with any available code. The flowchart of the ONIOM-EC-RISM/X scheme is shown in figure 4.6.

This scheme has the advantage that the individual EC-RISM calculations can be performed in parallel, thus reducing the total time required compared to the /A scheme, where the sub-calculations are performed strictly serially.

PMV-correction and extrapolation limit

Where the derivation of the PMV-correction for the other ONIOM-EC-RISM schemes was straightforward and similar to that of the *hr* approach, the nature of the /X approximation allows for multiple options to incorporate the correction into this scheme.

In principle, both the ONIOM extrapolation and the PMV correction must be considered ad hoc approximations, since neither can be derived directly from first principles. As a consequence, there is no intrinsic order to their composition, i.e. in principle one could either perform the ONIOM extrapolation first and then apply the PMV correction to the resulting expression, or first correct the individual free energies from the sub-calculations and then perform the ONIOM extrapolation. This distinction is not necessary for the other ONIOM-EC-RISM schemes, as the order of operation is predefined by performing the ONIOM extrapolation during the iterative solution of the solute free energy. In the following, these schemes are referred to as "ONIOM first" and "ONIOM second" or /Xa and /Xb, when a shorter notation is required, e.g. in the following equations.

The /Xa scheme where the ONIOM extrapolation is performed first is thus defined as

$$G_{\mathrm{sol}}^{/\mathrm{Xa,corr}} = G_{\mathrm{sol}}^{/\mathrm{X}} + c_V^{/\mathrm{Xa}}\Omega(\mathbf{V}_{\mathrm{m}}) + c_q^{/\mathrm{Xa}} q_{\mathrm{sol}} \tag{4.87}$$

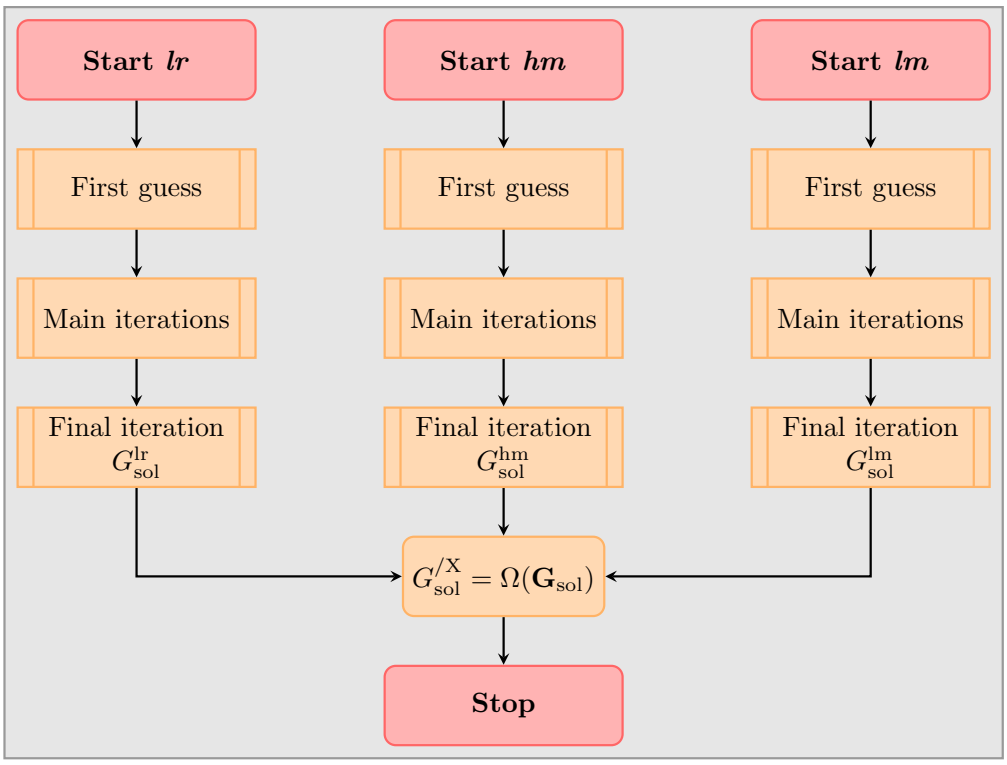

Figure 4.6.: Flowchart of the ONIOM-EC-RISM/X implementation. Each column represents an independent EC-RISM calculation using the respective level of theory and system.

and is similar to the previous corrections. Here, the extrapolated PMV $\Omega(\mathbf{V}_{\mathrm{m}}) = V^{\mathrm{lr}}_{\mathrm{m}} + V^{\mathrm{hm}}_{\mathrm{m}} - V^{\mathrm{lm}}_{\mathrm{m}}$ is used to keep this correction in line with the general idea of the /X scheme. $V^{\mathrm{lr}}_{\mathrm{m}}$, $V^{\mathrm{hm}}_{\mathrm{m}}$ and $V^{\mathrm{lm}}_{\mathrm{m}}$ denote partial molar volumes calculated from the ESP of the respective sub-calculation, i.e. $\varphi^{\mathrm{lr}}_{\mathrm{sol}}$, $\varphi^{\mathrm{hm}}_{\mathrm{sol}}$ and $\varphi^{\mathrm{lm}}_{\mathrm{sol}}$. Furthermore, this does not add to the computational cost as the individual volumes are available from the sub-calculations anyway, but it should be noted that in principle it would also be possible to use the *lr*-PMV $V^{\mathrm{lr}}_{\mathrm{m}}$.

If the ONIOM extrapolation is performed second, the /Xb expression

$$G^{/\mathrm{Xb,corr}}_{\mathrm{sol}} = \Omega(\mathbf{G}^{\mathrm{corr}}_{\mathrm{sol}}), \quad \mathbf{G}^{\mathrm{corr}}_{\mathrm{sol}} = \mathbf{G}_{\mathrm{sol}} + \begin{pmatrix} c^{\mathrm{lr}}_{V} V^{\mathrm{lr}}_{\mathrm{m}} \\ c^{\mathrm{hm}}_{V} V^{\mathrm{hm}}_{\mathrm{m}} \\ c^{\mathrm{lm}}_{V} V^{\mathrm{lm}}_{\mathrm{m}} \end{pmatrix} + \begin{pmatrix} c^{\mathrm{lr}}_{q} q_{\mathrm{sol}} \\ c^{\mathrm{hm}}_{q} q^{\mathrm{m}}_{\mathrm{sol}} \\ c^{\mathrm{lm}}_{q} q^{\mathrm{m}}_{\mathrm{sol}} \end{pmatrix} \tag{4.88}$$

is obtained, where in principle each solute free energy can be corrected with its own set of parameters. Here $q^{\mathrm{m}}_{\mathrm{sol}}$ denotes the charge of the *model* system. The last equation can

also be rewritten as

$$G_{\mathrm{sol}}^{/\mathrm{Xb,corr}} = G_{\mathrm{sol}}^{/\mathrm{X}} + \Omega \begin{pmatrix} c_V^{\mathrm{lr}} V_{\mathrm{m}}^{\mathrm{lr}} \\ c_V^{\mathrm{hm}} V_{\mathrm{m}}^{\mathrm{hm}} \\ c_V^{\mathrm{lm}} V_{\mathrm{m}}^{\mathrm{lm}} \end{pmatrix} + \Omega \begin{pmatrix} c_q^{\mathrm{lr}} q_{\mathrm{sol}} \\ c_q^{\mathrm{hm}} q_{\mathrm{sol}}^{\mathrm{m}} \\ c_q^{\mathrm{lm}} q_{\mathrm{sol}}^{\mathrm{m}} \end{pmatrix}, \tag{4.89}$$

which shows that the two correction schemes differ only in the order in which the PMV correction and the ONIOM extrapolation are evaluated for the correction terms. Therefore, the scheme shown in figure 4.6 is valid for both /Xa and /Xb. The two modes of correction are visually summarised in figure 4.7 and are identical if only one set of parameters is used for /Xb.

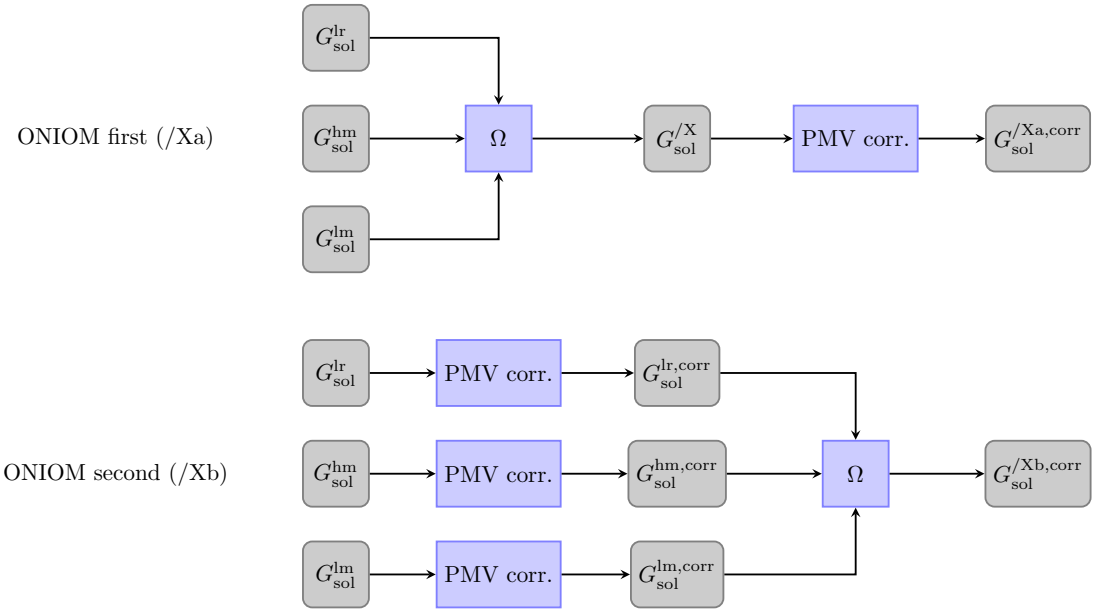

Figure 4.7.: Workflow of the two correction modes for ONIOM-EC-RISM/X. The thermodynamic quantities are shown in grey, while the ONIOM extrapolation Ω and the PMV correction are shown in blue. In the first mode, the ONIOM extrapolation is performed first and the PMV correction is applied to the resulting quantity. This order of operation is reversed in the second mode, thus requiring several sets of correction parameters.

As for the /A scheme, the PFL of the uncorrected /X scheme and the /Xa scheme from equation 4.87 gives the *hr* result. This is also true for /Xb if the parameters for *lr* and *lm* are identical.

In practice, all parameters used in this work are obtained by fitting partition-free models (such as *hr*, *lr* and /B@PFL) to experimental values, a process that will be explained in more detail in the next chapter. This avoids the need to partition the

molecules in the training dataset and removes the partition-dependence of the resulting parameters. Therefore, no unique sets of parameters can be obtained for the *model* subcalculations, *hm* and *lm*, and the parameters obtained from *hr* and *lr* need to be used. As a consequence of the respective identity of the *high* and *low* parameters the total number of parameter sets required for the /Xb correction mode is thus reduced from three to two and equation 4.89 can be rewritten as

$$G_{\mathrm{sol}}^{/\mathrm{Xb,corr}} = G_{\mathrm{sol}}^{/\mathrm{X}} + c_V^{\mathrm{lr}} \left(V_{\mathrm{m}}^{\mathrm{lr}} - V_{\mathrm{m}}^{\mathrm{lm}} \right) + c_V^{\mathrm{hr}} V_{\mathrm{m}}^{\mathrm{hm}} + c_q^{\mathrm{lr}} \left(q_{\mathrm{sol}} - q_{\mathrm{sol}}^{\mathrm{m}} \right) + c_q^{\mathrm{hr}} q^{\mathrm{m}}. \tag{4.90}$$

In case the partitioning scheme results in $q_{\mathrm{sol}} = q_{\mathrm{sol}}^{\mathrm{m}}$, the charge correction terms reduce to $c_q^{\mathrm{hr}} q_{\mathrm{sol}}$.

Part III.

Calculation of molecular properties with ONIOM-EC-RISM

5. Validation of ONIOM-EC-RISM on the SAMPL6 data set

In the following chapters the ONIOM-EC-RISM multiscale solvation model will be used to calculate the acidity constants of molecules of increasing size. This serves several purposes. Firstly, to validate the solvation model by comparison with the results obtained from the extrapolation target. Secondly, it will be shown how the ONIOM-EC-RISM method can be applied to systems that are too large to be modelled with the *hr*-EC-RISM approach, and in particular what considerations need to be made in order to transfer previous workflows to the context of this new solvation model. Furthermore, the accurate calculation of acidity constants is a considerable challenge and therefore a great opportunity to access the accuracy of solvation models.

The first part of this endeavour, the initial method validation, will be carried out by predicting pK_a values for the small molecule data set from the SAMPL6 blind prediction challenge.[25] It has already been shown that EC-RISM can be used to obtain accurate estimates of pK_a values for the SAMPL6 data set.[24] This was done by training a PMV correction by fitting the model to experimental solvation free energies from the Minnesota Solvation Database (MNSOL).[117] In a second step, a correction for the calculated pK_a values was similarly trained by fitting a linear model to a data set of experimental values from Klicić et al.[118] The resulting model was then used to blindly predict pK_a values during the SAMPL6 challenge. This process is depicted in figure 5.1.

In order to calculate pK_a values in an analogue manner in this work, this workflow has to be adapted for the ONIOM-EC-RISM model. The first part of this problem, the training of the PMV correction, is addressed in the following section.

5.1. Parameterisation of the PMV correction

Fitting the two-parameter model of the PMV correction requires the initial generation of structures as shown in figure 5.1. In their SAMPL6 publication[24], which uses *hr*-EC-RISM and serves as an reference throughout this chapter, Tielker et al. used a RDKit based conformer search algorithm to generate these structures.[23,24] An initial geometry optimisation of these structures was then performed with the GAFF 1.5 force field, AM1-BCC charges, followed by a clustering step with a 5 kcal mol^{-1} energy window to minimise the total number of structures. In a subsequent step, geometry optimisations were performed by the authors at the B3LYP/6-311+G** level of theory with a PCM

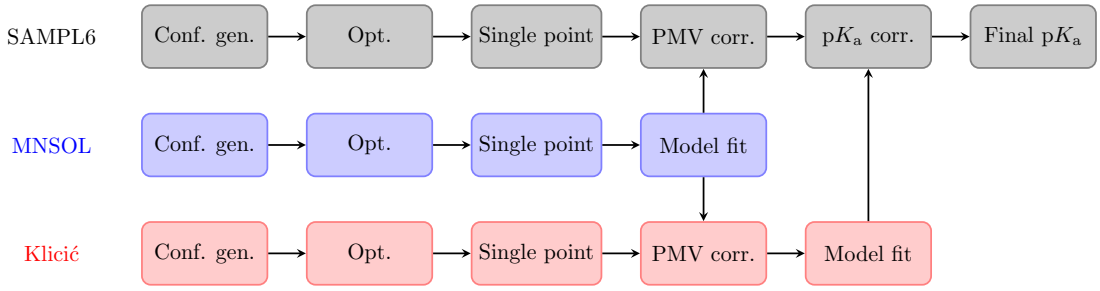

Figure 5.1.: Workflow for the SAMPL6 pK_a blind prediction challenge, based on EC-RISM single point calculations. The final model consists of two empirical corrections: A PMV correction parameterised to experimental solvation free energies from the MNSOL database[117] and a pK_a correction obtained similarly by fitting a linear model to experimental pK_a values from the Klicić dataset.[118] This workflow needs to be transferred to the ONIOM-EC-RISM framework.

water solvent in Gaussian09. Only the global energetically minimal structures were used for the subsequent EC-RISM calculations. Vacuum conformers were generated by re-optimising the solvated structures at the corresponding level of theory without the PCM model. From these vacuum structures the global minimal conformer was selected. Single point energies were obtained by *hr*-EC-RISM and vacuum calculations at the MP2/6-311+G** level of theory.[24]

The free energies of solvation

$$\Delta_{\mathrm{solv}} G^{0,\mathrm{hr}}_{\mathrm{calc}}(c_V^{\mathrm{hr}}, c_q^{\mathrm{hr}}) = G^{\mathrm{hr,corr}}_{\mathrm{sol}}(c_V^{\mathrm{hr}}, c_q^{\mathrm{hr}}) - E^{\mathrm{hr}}_{\mathrm{vac}} \tag{5.1}$$

calculated as a function of the PMV correction parameters c_V^{hr} and c_q^{hr} were then fitted by Tielker et al. to the experimental free energies of solvation from the water subset of the MNSOL database by minimising the expression

$$\{c_V^{\mathrm{hr}}, c_q^{\mathrm{hr}}\} = \underset{c_V^{\mathrm{hr}}, c_q^{\mathrm{hr}}}{\arg\min} \left[\sum_j \left(\Delta_{\mathrm{solv}} G^{0,\mathrm{hr}}_{\mathrm{calc},j}(c_V^{\mathrm{hr}}, c_q^{\mathrm{hr}}) - \Delta_{\mathrm{solv}} G^{0}_{\mathrm{exp},j} \right)^2 \right], \tag{5.2}$$

where the sum runs over all molecules j from the subset.

5.1.1. Partition-free PMV corrections

The adaptation of this parameterisation strategy for ONIOM-EC-RISM calculations therefore requires an analogue of the conformer generation workflow described above, as well as the replacement of the *hr*-free solvation energies by the respective ONIOM quantity. The latter problem will be discussed first, as it induces a solution to the set of geometries required for the ONIOM-PMV correction.

During the discussion of the size extrapolation limit of the ONIOM-EC-RISM schemes in section 4.6, it was already briefly outlined that the size of the training data set leads to significant problems for the parameterisation of an ONIOM-PMV correction. The water subset of the MNSOL contains over 500 entries. A parameterisation using the standard ONIOM-EC-RISM approach would require the partitioning of all molecules in the dataset, which for the implementation presented here implies the specification of the *model* system subset via their indices. Doing this manually is not feasible and would introduce an arbitrary partitioning error into the final prediction quality. Furthermore, there is currently no algorithm that can predict the optimal set of partitions that minimises this error. In addition, the MNSOL dataset consists of a number of small molecules containing only a few atoms, such as methane, which do not allow for chemically meaningful partitions. It is therefore neither realisable in a reasonable time nor is it feasible to parameterise a PMV correction for the ONIOM-EC-RISM schemes on a data set of partitioned molecules.

To avoid these problems, the ONIOM PMV corrections are parameterised using the size extrapolation limit of the ONIOM-EC-RISM schemes. By applying the PFL to the PMV-corrected free energy of the system, equations are obtained that do not require a partition to be specified for the solute, as outlined in section 4.6.

Since for /A the PMV correction at the PFL does not differ from the *hr*-PMV correction, it is reasonable to apply the *hr*-parameters, thus no new parameterisation is required and the parameters from the Tielker et al. publication[24] can be reused. This parameterisation strategy may be motivated by the fact that if the ONIOM extrapolation is able to accurately reproduce the *hr* target, then the application of the *hr* parameterisation is also a consistent parameterisation for the /A scheme.

In contrast, the PMV correction for the ONIOM-EC-RISM/B@PFL scheme is a mixed scheme where the energy of the *real* system is evaluated with the *high* level theory in a solvent environment calculated from the *lr*-ESP. The parameterisation of the correction at this level of theory, therefore, gives rise to a set of parameters that differ from the *hr*-parameters, if the *lr*-ESP is unable to reproduce the *hr*-ESP. If the contrary is true, it would be reasonable to also use the *hr*-parameters to approximate the /B parameters that would be obtained from calculations on partitioned molecules. See the previous chapter for a more detailed discussion. It is therefore reasonable to test both the /B@PFL and *hr* PMV parameters in conjunction with the /B scheme. The later is already available from the SAMPL6 reference publication by Tielker et al.[24] and the former will be derived in the following sections.

The two correction modes of the /X scheme provide multiple options for the PMV correction. The first mode, /Xa, where the ONIOM extrapolation is performed first, is identical to *hr* at the partition free limit. As with the /A scheme, the use of the *hr* parameters is therefore appropriate.

In the second mode, /Xb, the ONIOM extrapolation is performed after the correction of the individual solute free energies. Consequently, this mode allows the use of three

independent correction parameters. Both *lr* and *lm* use the same level of theory, so it is reasonable to generate a set of PMV correction parameters at the *low* level of theory and apply them to both sub-calculations. Consequently, in addition to the *hr* parameterisation, an *lr* parameterisation is required for this correction mode. Note that if only one set of parameters is used, both modes are identical. The first mode can thus be considered as a special case of the second. Again, see the previous chapter for more details. As the *hr*-parameters are already known, only the *lr*parameters need to be derived in the following sections.

Due to the limited advantages over the /B scheme as explained in section 4.6.3, the /C scheme was discarded and will not be considered for PMV corrections or pK_a calculations.

In summary, in addition to the already existing *hr*-PMV correction, parameters for the /B@PFL scheme as well as *lr* parameters are required for the consistent construction of the different ONIOM-PMV corrections.

However, it should be emphasised that the use of the PFL is only a way of circumventing the partitioning limitations imposed by the size of the training data set. The true parameters that would be obtained by training the ONIOM schemes on partitioned molecules are not available, so the PFL parameterisations must be seen as an approximation to these parameters. A large part of this chapter is therefore devoted to testing the validity of this approximation through the prediction of solvation free energies and acidity constants and direct comparison with experimental reference data.

In the case of the /B scheme, the calculation of the solvation free energies also requires the calculation of the vacuum energy via an ONIOM expression giving

$$\Delta_\mathrm{solv} G^{0,/\mathrm{B}}_\mathrm{calc}(c_V^{/\mathrm{B}}, c_q^{/\mathrm{B}}) = G^{/\mathrm{B,corr}}_\mathrm{sol}(c_V^{/\mathrm{B}}, c_q^{/\mathrm{B}}) - \Omega(\mathbf{E}_\mathrm{vac}). \tag{5.3}$$

The application of the PFL thus results in the equation

$$\Delta_\mathrm{solv} G^{0,/\mathrm{B@PFL}}_\mathrm{calc}(c_V^{/\mathrm{B@PFL}}, c_q^{/\mathrm{B@PFL}}) = G^{/\mathrm{B@PFL,corr}}_\mathrm{sol}(c_V^{/\mathrm{B@PFL}}, c_q^{/\mathrm{B@PFL}}) - E^\mathrm{hr}_\mathrm{vac}, \tag{5.4}$$

which can be parameterised using the approach for the *hr* correction from equation 5.2. The *lr* parameters are also derived in an analogous manner.

The *hr* parameterisation workflow chosen by Tielker et al. uses a PCM solvent for geometry optimisation as a proxy for the EC-RISM potential surface, as there is currently no implementation that allows geometry optimisation with EC-RISM. Of course, for the /B@PFL parameterisation this would correspond to an optimisation with an ONIOM-PCM solvent. Gaussian offers ONIOM-PCM implementations for /A, /B, /C and /X, but no optimisation is available for the /B scheme. Therefore the *hr*-PCM geometries will be used in the following. This can be justified as the /B scheme attempts to extrapolate to the corresponding *hr* energy surface and the /B@PFL is identical to *hr* if the *lr*-ESP is able to accurately reproduce the *hr*-ESP. To additionally estimate the effect

of a lower cost optimisation level, the *lr* geometries are also used to generate a PMV parameterisation.

These two sets of geometries are also used for the *lr*-parameterisation, so that it is possible to additionally measure the effect of the *hr*-ESP in contrast to the *lr*-ESP, and also the effect of the *hr*-description of the solute in contrast to its *lr*-description.

Table 5.1 provides an overview of the PMV correction models, as well as an identifier that is also used in the tables showing the parameterisation and prediction results. The two net charge correction modes shown in this table are motivated and explained in section 5.1.3.

The model parameters can be quickly retrieved using the links to the corresponding tables given in table 5.1. The model IDs specify from left to right: the level of theory (/B@PFL or *lr*), the set of geometries (B: B3LYP-PCM or P: PM6-PCM), the ESP-approximation (M: multipole, N: NDDO or P: point-charge) and the two net charge correction modes, i.e. whether one set of parameters was used for all charge states (A: all) or individual corrections were applied (I: indiv.).

5.1.2. Computational details

The structures were reoptimised starting from the PCM and vacuum optimised structures used in the original SAMPL6 publications. These structures were provided by Nicolas Tielker. Reoptimisations were performed at the PM6 level of theory with Gaussian16[42] in vacuum or using an IEFPCM water model. Optimisations were repeated with the "CalcFC" keyword if the calculation failed. If this did not result in convergence, the optimisations were repeated using the "CalcAll" keyword. If convergence was still not achieved, the structure was discarded. The corresponding data can be found in the electronic supplementary material.[116]

Single point calculations in solution were performed with the *lr*-EC-RISM and ONIOM-EC-RISM/B@PFL schemes, using MP2/6-311+G** for the *high* theory level and PM6 for the *low* theory level. Point charge and exact potential models were used to investigate the effect of the solute potential supplied to the 3D-RISM solver. Calculations with the exact potential were performed using the "Mult" or "NDDO" option in EMPIRE's eh5cube tool. The first option provides a multipole based ESP approximation while the second provides an ESP calculated directly from the NDDO wavefunction using Kloppmann-Ohno parameters, as stated in the eh5cube options.

All EC-RISM calculations were performed with the PSE-2 closure on a 128^3 grid with 0.3 Å grid spacing in all three dimensions at a temperature of 298.15 K. The calculations were converged to a threshold of $0.01\,\mathrm{kcal\,mol^{-1}}$. All 3D-RISM settings are identical to those used in the original SAMPL6 publication by Tielker et al.[24] Lennard Jones parameters were taken from GAFF 1.5.[119] Pre-computed solvent susceptibility functions obtained from DRISM using SPC/E parameters[120] and the experimental isothermal water compressibility of $0.450\,183 \cdot 10^{-9}\,\mathrm{Pa^{-1}}$ were used. As in the SAMPL5[23] and

Table 5.1.: Overview of all PMV correction models parameterised in this chapter. In addition, a model identifier is given in the last column. See section 5.1.3 for an explanation of the two net charge correction modes.

EC-RISM model	Optimisation	Potential	Charge fit	Model ID	Table
lr	PM6-PCM	Exact (Mult.)	All	*lr*|P|M|A	(5.2)
			Indiv.	*lr*|P|M|I	(5.2)
		Exact (NDDO)	All	*lr*|P|N|A	(5.2)
			Indiv.	*lr*|P|N|I	(5.2)
		Point charge	All	*lr*|P|P|A	(5.2)
			Indiv.	*lr*|P|P|I	(5.2)
	B3LYP-PCM	Exact (Mult.)	All	*lr*|B|M|A	(5.3)
			Indiv.	*lr*|B|M|I	(5.3)
		Exact (NDDO)	All	*lr*|B|N|A	(5.3)
			Indiv.	*lr*|B|N|I	(5.3)
		Point charge	All	*lr*|B|P|A	(5.3)
			Indiv.	*lr*|B|P|I	(5.3)
/B@PFL	PM6-PCM	Exact (Mult.)	All	/B|P|M|A	(5.4)
			Indiv.	/B|P|M|I	(5.4)
		Exact (NDDO)	All	/B|P|N|A	(5.4)
			Indiv.	/B|P|N|I	(5.4)
		Point charge	All	/B|P|P|A	(5.4)
			Indiv.	/B|P|P|I	(5.4)
	B3LYP-PCM	Exact (Mult.)	All	/B|B|M|A	(5.5)
			Indiv.	/B|B|M|I	(5.5)
		Exact (NDDO)	All	/B|B|N|A	(5.5)
			Indiv.	/B|B|N|I	(5.5)
		Point charge	All	/B|B|P|A	(5.5)
			Indiv.	/B|B|P|I	(5.5)

SAMPL6[24] publications, partial molar volumes were calculated using the direct correlation function route.

ORCA calculations were started with the "TightSCF" convergence criteria. The convergence criteria in EMPIRE were set to $10^{-6}\,\mathrm{kcal\,mol^{-1}}$ for energy convergence and 10^{-6} for convergence of the maximum off-diagonal CFC element for ONIOM-EC-RISM calculations and to $10^{-4}\,\mathrm{kcal\,mol^{-1}}$ and 10^{-4}, respectively for *lr*-EC-RISM calculations. However, it is still an order of magnitude smaller than the default defined in EMPIRE and it has been confirmed that the different convergence criteria lead to nearly identical results for the solvent distribution by comparing the resulting excess chemical potentials. The associated data and statistical parameters are shown in figure 1 and table 1 in the appendix.

The corresponding vacuum calculations were performed using EMPIRE and ORCA respectively, using the convergence criteria mentioned above.

Only the globally optimal structures obtained from the gas phase and PCM optimisations respectively were used to parameterise the PMV correction. This is identical to the procedure described in the original SAMPL6 publication.[24]

To ensure identical parameterisation with respect to previous models, the Mathematica notebooks used for the SAMPL6 publication were provided by Nicolas Tielker and only modified to allow the use of the ONIOM-EC-RISM input. The associated raw data, such as atomic coordinate files, can be found in the electronic supplementary material.[116]

5.1.3. Solvation free energy prediction with PM6-EC-RISM

PM6 reoptimised structures

Figure 5.2 and table 5.2 show the prediction results for the free energy of solvation on the MNSOL water subset with the *lr*-EC-RISM solvation model, or more precisely the PM6-EC-RISM model. To investigate the prediction quality of the different models, a descriptive regression was performed, resulting in the slope parameter m' and the intercept parameter b'. Optimal models result in a slope of one and an intercept of zero.

Note that the parameters m' and b' are obtained by fitting $\Delta_{\mathrm{solv}}G^0_{\mathrm{calc}} = m'\Delta_{\mathrm{solv}}G^0_{\mathrm{exp}} + b'$. The former was chosen over the commonly applied $\Delta_{\mathrm{solv}}G^0_{\mathrm{exp}} = m''\Delta_{\mathrm{calc}}G^0_{\mathrm{exp}} + b''$ approach, as it is the method used in the SAMPL6 paper by Tielker et al.,[24] which serves as a reference study for this work. Therefore, the same procedure is applied here. In particular, this allows for comparison of the newly parameterised models with the *hr*-reference models.

In addition, the root mean square error (RMSE), mean absolute error (MAE) and mean signed error (MSE) were calculated for the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,\mathrm{lr}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$. For comparison, the best performing models from the original SAMPL6 publication by Tielker et al.[24] are reported alongside the new models. Note that these

original models were all parameterised on structures optimised at the B3LYP/6-311+G** level of theory with PCM, but are also shown in tables where PM6-PCM geometries were used to give a better overview of the effect of the more approximate level of theory.

Using the multipole approximation of the exact potential (*lr*|P|M|A) results in an overall RMSE of 5.74 kcal mol^{-1}. The associated correction parameters are very different from the corresponding *hr* parameters. The charge correction parameter c_q is much smaller than the *hr* value with a value of $-26.236\,\mathrm{kcal\,mol^{-1}}\,e^{-1}$, while the pressure parameter c_V is larger than the reference value. Comparing the results for the multipole ESP with the point charge potential (*lr*|P|P|A) shows that the correction parameters are almost identical and the statistical quantities are very similar. The point charge model uses a potential that is the sum of all point charge potentials derived from Coulson atomic charges, and thus represents a monopole approximation to the ESP. It is therefore reasonable to expect that the addition of higher order terms to the ESP will lead to a significant improvement in the prediction of solvation free energies. In fact, the multipole ESP gives a higher RMSE than the corresponding point-charge model.

Calculating the electrostatic potential (ESP) from the "Neglect of Diatomic Differential Overlap" (NDDO) wave function (*lr*|P|N|A) leads to an improved RMSE of 4.40 kcal mol^{-1}, which is more than 1 kcal mol^{-1} better than what is obtained from the multipole-based ESP.

It is noticeable across all models that the prediction for neutral species gives similar results with RMSE values between 3.16 and 3.70 kcal mol^{-1}. However, significantly poorer outcomes are observed for ionic species. Especially anionic molecules yield a RMSE considerably higher than 10 kcal mol^{-1} for the point charge and multipole-based model. This aligns with the findings of other researchers who employed PM6 with the SMD and COSMO solvation models to predict solvation free energies on the MNSOL dataset.[121] In the case of the SMD solvation model, the authors could enhance the predictive outcomes by modifying the radii parameter of a particular group of atoms. As RISM does not specify any explicit cavity, this procedure cannot be easily transferred into the *lr*-EC-RISM model. Therefore, an alternative approach must be identified to rectify the errors related to ions.

From figure 5.2, it is evident that the use of a single set of correction parameters cannot accurately compensate for the errors observed for ionic species. Therefore, additional PMV corrections were parameterised where each charge state was corrected by fitting an individual linear model. Thus three sets of correction parameters are obtained. The corresponding results are also depicted in figure 5.2 and table 5.2.

This alternative method of fitting produces an improvement in the RMSE for the statistic for all charge states collectively and individual states specifically. In particular, the models for neutral and cationic molecules, which have both multipole (*lr*|P|M|I) and point charge ESP (*lr*|P|P|I), generate more accurate predictions compared to those produced with only one parameter set. Moreover, there has been enhancement in the prediction quality for anions, although considerable deviations with an RMSE over

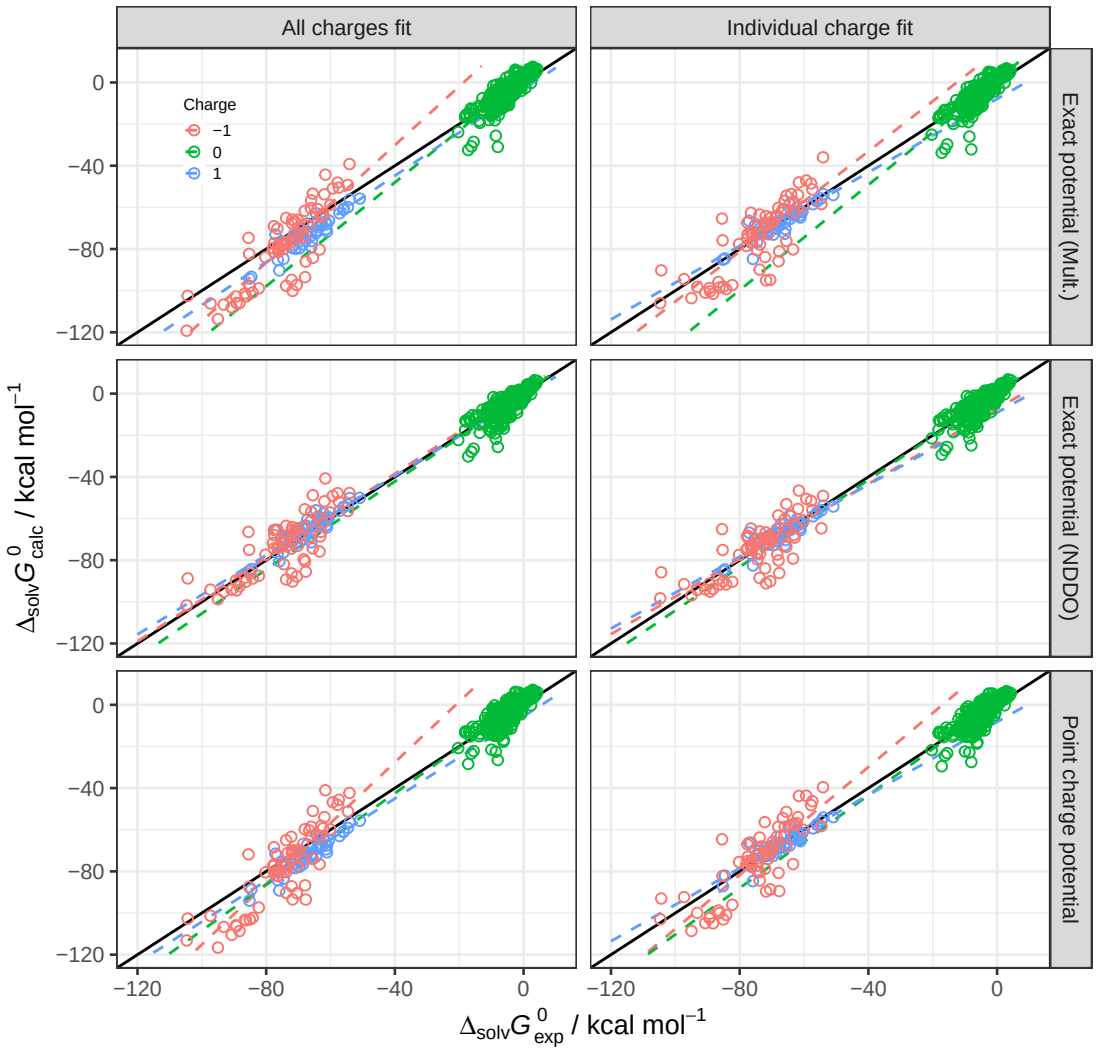

Figure 5.2.: Predicted versus experimental free energies of solvation for calculations at the PM6-EC-RISM level of theory on the reoptimised PM6-PCM conformers. The first column shows predictions made with one set of PMV correction parameters for all charge states, while in the second column each state has been parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Anionic, neutral and cationic species are shown in red, green and blue, respectively. The corresponding numerical model data are shown in table 5.2.

Table 5.2.: Statistical quantities and PMV correction parameters obtained for the free energy of solvation prediction on the MNSOL water subset. Single-point calculations were with PM6-EC-RISM on structures reoptimised with PM6-PCM. The statistical quantities RMSE, MAE and MSE have been calculated for the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,\mathrm{lr}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$ and are given in kcal mol^{-1}. The parameters m' and b' were obtained by descriptive regression, where the latter is also given in kcal mol^{-1}. The correction parameters c_V and c_q are given in kcal mol^{-1} Å^{-3} and kcal mol^{-1} e^{-1}. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. The corresponding plots are shown in figure 5.2.

Potential & ID	Charge fit	q^{lr}	RMSE	MAE	MSE	m'	b'	R^2	c_V	c_q
Exact (Mult.)	All	All	5.74	3.88	-0.56	1.09	1.29	0.98	-0.0795	-26.236
lr|P|M|A		-1	11.68	9.15	-4.16	1.41	26.26	0.72		
		0	3.58	2.66	0.75	1.25	1.86	0.73		
		1	6.78	6.04	-5.81	1.04	-3.32	0.82		
Exact (Mult.)	Individual	All	4.91	3.26	0.15	1.01	0.37	0.97		
lr|P|M|I		-1	9.99	8.19	0.00	1.20	15.05	0.66	-0.1773	-36.943
		0	3.53	2.51	0.20	1.26	1.37	0.73	-0.0826	1.000
		1	3.17	2.21	-0.00	0.88	-7.89	0.81	-0.1076	-15.889
Exact (NDDO)	All	All	4.40	2.92	0.25	0.98	-0.20	0.98	-0.0818	-21.148
lr|P|N|A		-1	8.73	6.98	1.30	1.01	1.79	0.63		
		0	3.16	2.21	-0.13	1.06	0.12	0.68		
		1	3.69	2.82	1.81	0.95	-1.42	0.82		
Exact (NDDO)	Individual	All	4.22	2.84	0.17	1.00	0.19	0.98		
lr|P|N|I		-1	8.29	6.39	0.00	0.90	-7.29	0.60	-0.1364	-23.753
		0	3.13	2.28	0.22	1.05	0.43	0.68	-0.0798	1.000
		1	3.10	2.30	0.00	0.86	-9.13	0.82	-0.0984	-20.303
Point charge	All	All	5.54	3.92	-0.11	1.08	1.51	0.97	-0.0807	-26.050
lr|P|P|A		-1	11.20	8.75	-3.33	1.46	30.65	0.75		
		0	3.70	2.93	1.05	1.10	1.49	0.65		
		1	5.66	4.98	-4.65	0.99	-5.55	0.83		
Point charge	Individual	All	5.02	3.48	0.45	1.02	0.76	0.97		
lr|P|P|I		-1	10.14	8.21	0.00	1.30	21.99	0.70	-0.1557	-34.527
		0	3.66	2.80	0.58	1.11	1.09	0.65	-0.0832	1.000
		1	3.07	2.24	0.00	0.88	-8.24	0.82	-0.1008	-18.131
Exact	All	All	2.04	1.43	-0.26	1.00	-0.35	1.00	-0.1025	-15.728
(*hr*-ref.)[24]		-1	3.07	2.46	0.01	1.10	7.18	0.94		
hr|B|E|A		0	1.56	1.13	-0.36	0.97	-0.47	0.89		
		1	2.98	2.10	0.02	0.96	-2.62	0.85		
Point charge	All	All	2.98	2.01	-0.56	1.04	0.42	0.99	-0.1009	-20.542
(*hr*-ref.)[24]		-1	5.27	3.92	-2.23	1.18	10.91	0.88		
hr|B|P|A		0	1.77	1.31	0.20	1.04	0.36	0.87		
		1	4.66	4.16	-3.61	0.94	-7.47	0.85		

8.00 kcal mol^{-1} can still be observed. Overall the NDDO-ESP model (lr|P|N|I) shows the smallest improvements for this level of theory.

B3LYP optimised structures

Performing identical predictions on the geometries from the original SAMPL6 publication generates the outcomes presented in table 5.3 and figure 5.3.

Comparison of these predictions reveals the effect of optimising with the higher level of theory B3LYP compared to the less costly PM6 Hamiltonian. These improved geometries lead to a slightly better prediction quality for the statistic across all charge states. Specifically, the RMSE for both multipole models (lr|B|M|A and lr|B|M|I) can be improved from 5.74 to 5.63 kcal mol^{-1}, and from 4.91 to 4.77 kcal mol^{-1}, respectively.

Similar small improvements can be observed in the statistics for each individual charge state. Thus it can be argued that the suboptimal predictive performance primarily stems from deficiencies in the solute description at the PM6 level and the usage of the PM6-ESP for the solvent structure calculation, or a combination of these sources of errors, rather than being a result of the PM6-PCM structure reoptimisation.

5.1.4. Solvation free energy prediction with ONIOM-EC-RISM/B@PFL

PM6 reoptimised structures

The ONIOM-EC-RISM/B@PFL scheme can be considered as a model that represents a midpoint between the lr-EC-RISM and hr-EC-RISM models, as it uses a hr description of the solute and a lr description of the solvent environment. This model therefore allows the isolated effect of the improved solute description to be estimated at the *high* level of theory, while retaining the same solvent description as for the lr-EC-RISM calculations. Similarly, the difference in prediction quality between the /B@PFL and hr models can be attributed to the use of the expensive *high* level method for calculating the solvent environment and can therefore be used to investigate the change in ESP while keeping the solute description constant.

The results for the /B@PFL PMV correction schemes for the PM6-PCM structures are shown in figure 5.4 and table 5.4. Statistics were calculated from the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,/\mathrm{B@PFL}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$.

Compared to previous prediction results, the improved description of the solute by the *high* level method leads to significantly better results. For the multipole model, "/B|P|M|A", where one set of parameters is used for all charge states, an RMSE of 3.75 kcal mol^{-1} is obtained. The second correction with individual fits for each charge state (/B|P|M|I) further improves the results to 3.40 kcal mol^{-1}. The corresponding point charge models (/B|P|P|A and /B|P|P|I) give similar correction parameters and also similar statistical quantities. For the NDDO-ESP model with one set of parameters for all charge states (/B|P|N|A) also improved results can be observed, but the change in

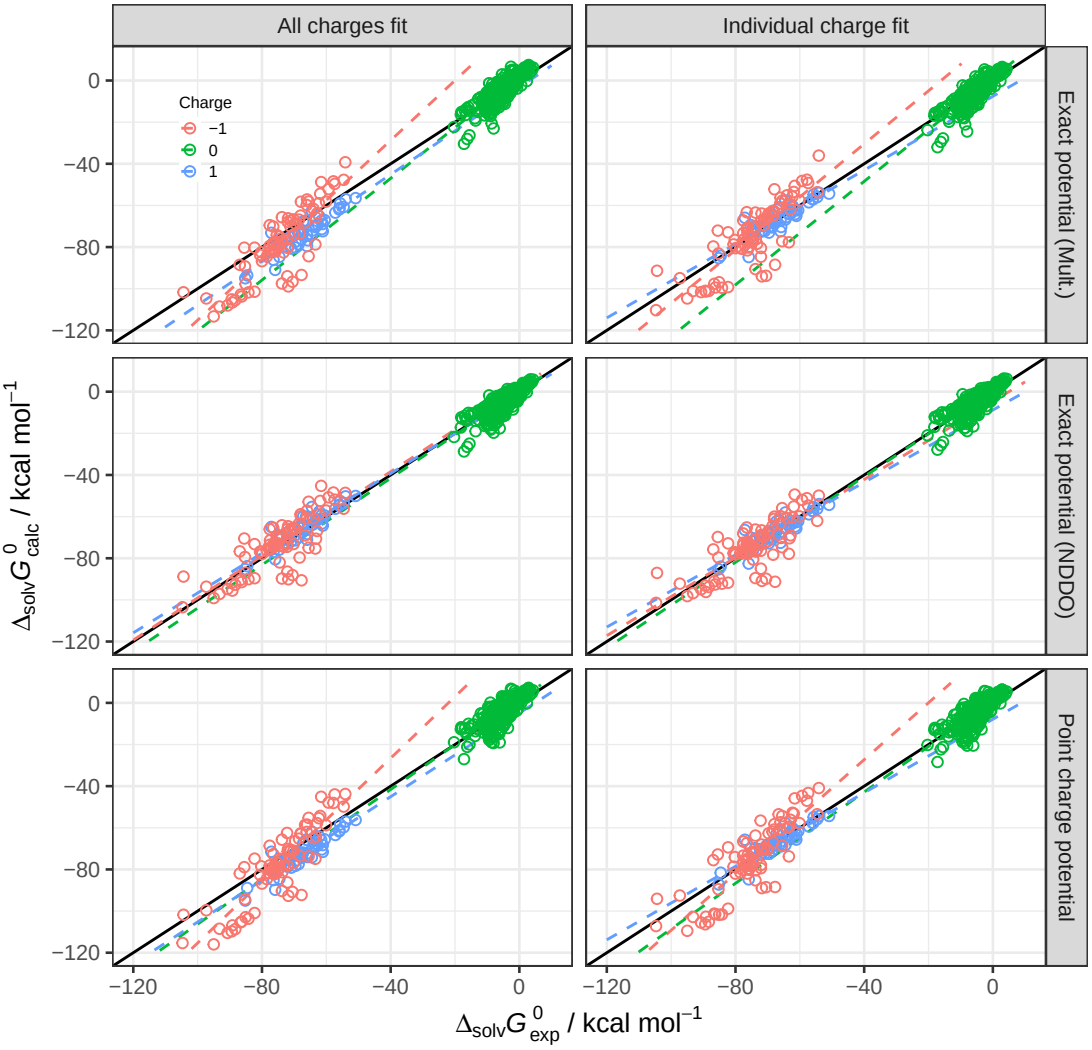

Figure 5.3.: Predicted versus experimental free energies of solvation for calculations at the PM6-EC-RISM level of theory on the PCM conformers optimised at the B3LYP/6-311+G** level of theory. The first column shows predictions made with one set of PMV correction parameters for all charge states, while in the second column each state was parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Anionic, neutral and cationic species are shown in red, green and blue, respectively. The corresponding numerical model data are shown in table 5.3.

Table 5.3.: Statistical quantities and PMV correction parameters obtained for the free energy of solvation prediction on the MNSOL water subset. Single-point calculations were performed at the PM6-EC-RISM level of theory on the B3LYP/6-311+G**-PCM structures. The statistical quantities RMSE, MAE and MSE have been calculated for the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,\mathrm{lr}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$ and are given in kcal mol^{-1}. The parameters m' and b' were obtained by descriptive regression, where the latter is also given in kcal mol^{-1}. The correction parameters c_V and c_q are given in kcal mol^{-1} Å^{-3} and kcal mol^{-1} e^{-1}. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. The corresponding plots are shown in figure 5.3.

Potential & ID	Charge fit	q^{lr}	RMSE	MAE	MSE	m'	b'	R^2	c_V	c_q
Exact (Mult.)	All	All	5.63	3.92	-0.63	1.09	1.39	0.98	-0.0802	-26.988
lr|B|M|A		-1	10.69	8.18	-4.03	1.44	28.76	0.75		
		0	3.38	2.67	0.85	1.22	1.84	0.74		
		1	7.31	6.60	-6.37	1.05	-3.30	0.82		
Exact (Mult.)	Individual	All	4.77	3.27	0.12	1.01	0.35	0.98		
lr|B|M|I		-1	9.33	7.67	-0.00	1.27	20.53	0.69	-0.1545	-36.167
		0	3.29	2.48	0.16	1.24	1.24	0.75	-0.0839	1.000
		1	3.26	2.29	-0.00	0.89	-7.58	0.81	-0.1091	-15.942
Exact (NDDO)	All	All	4.21	2.90	0.26	0.98	-0.22	0.98	-0.0829	-21.094
lr|B|N|A		-1	7.86	6.31	1.20	1.01	2.00	0.65		
		0	3.00	2.19	-0.16	1.04	0.02	0.70		
		1	3.74	2.83	1.89	0.96	-1.01	0.83		
Exact (NDDO)	Individual	All	4.07	2.81	0.15	1.00	0.18	0.98		
lr|B|N|I		-1	7.60	5.80	0.00	0.94	-4.67	0.63	-0.1196	-22.622
		0	2.97	2.25	0.20	1.03	0.33	0.70	-0.0809	1.000
		1	3.12	2.31	0.00	0.87	-8.80	0.82	-0.0994	-20.334
Point charge	All	All	5.52	4.01	-0.16	1.08	1.62	0.98	-0.0808	-26.935
lr|B|P|A		-1	10.35	8.09	-3.26	1.48	32.61	0.77		
		0	3.66	2.96	1.15	1.08	1.50	0.65		
		1	6.17	5.50	-5.15	1.00	-4.93	0.82		
Point charge	Individual	All	5.00	3.52	0.42	1.02	0.76	0.97		
lr|B|P|I		-1	9.55	7.71	-0.00	1.37	27.39	0.73	-0.1316	-33.813
		0	3.61	2.80	0.56	1.10	0.99	0.65	-0.0840	1.000
		1	3.21	2.35	0.00	0.88	-7.85	0.81	-0.1030	-18.176
Exact	All	All	2.04	1.43	-0.26	1.00	-0.35	1.00	-0.1025	-15.728
(*hr*-ref.)[24]		-1	3.07	2.46	0.01	1.10	7.18	0.94		
hr|B|E|A		0	1.56	1.13	-0.36	0.97	-0.47	0.89		
		1	2.98	2.10	0.02	0.96	-2.62	0.85		
Point charge	All	All	2.98	2.01	-0.56	1.04	0.42	0.99	-0.1009	-20.542
(*hr*-ref.)[24]		-1	5.27	3.92	-2.23	1.18	10.91	0.88		
hr|B|P|A		0	1.77	1.31	0.20	1.04	0.36	0.87		
		1	4.66	4.16	-3.61	0.94	-7.47	0.85		

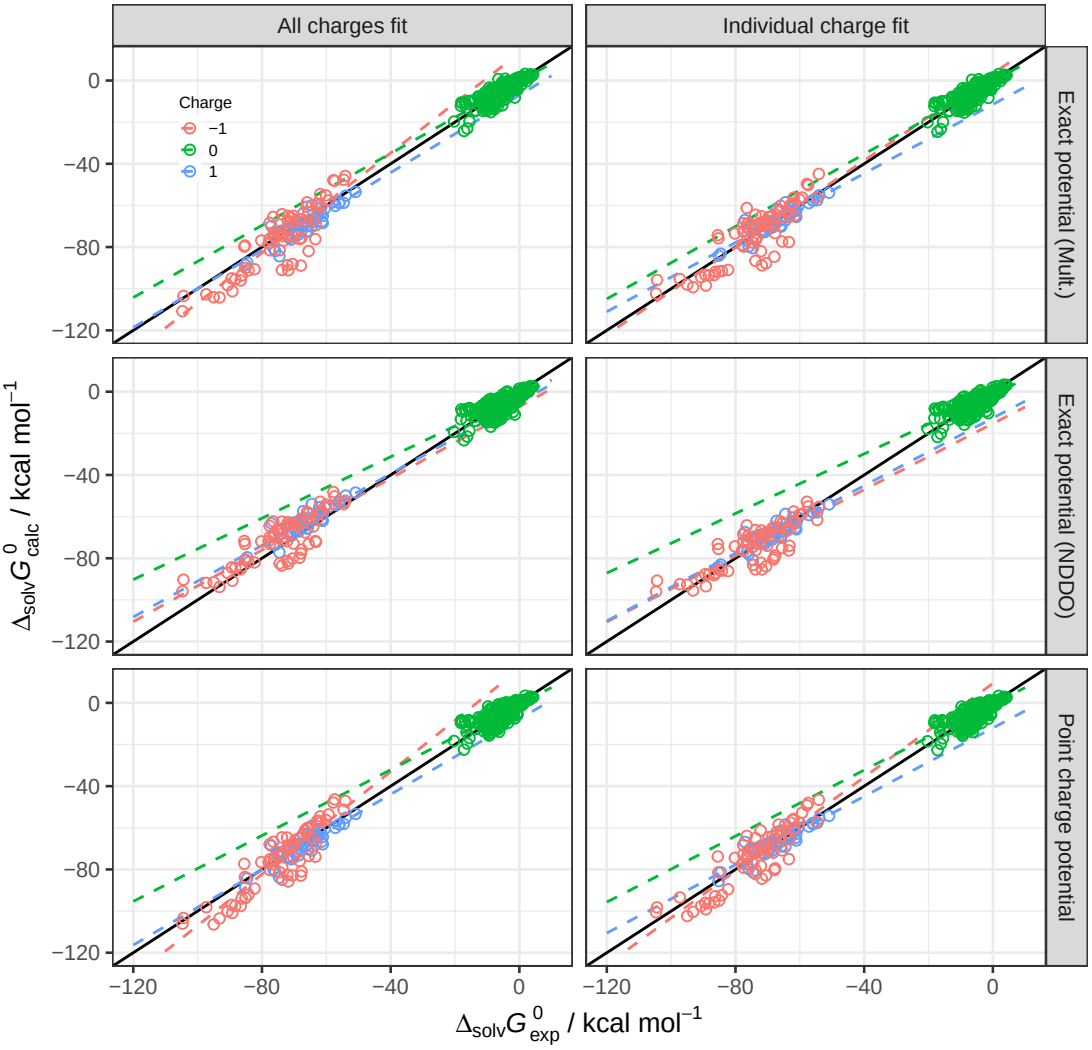

Figure 5.4.: Predicted versus experimental free energies of solvation for calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on the reoptimised PM6-PCM conformers. The first column shows predictions made with one set of PMV correction parameters for all charge states, while in the second column each state has been parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Anionic, neutral and cationic species are shown in red, green and blue, respectively. The corresponding numerical model data are shown in table 5.4.

Table 5.4.: Statistical quantities and PMV correction parameters obtained for the free energy of solvation prediction on the MNSOL water subset. Single-point calculations were performed at the ONIOM2(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on structures reoptimised with PM6-PCM. The statistical quantities RMSE, MAE and MSE have been calculated for the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,/\mathrm{B@PFL}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$ and are given in kcal mol^{-1}. The parameters m' and b' were obtained by descriptive regression. The correction parameters c_V and c_q are given in kcal mol^{-1} Å^{-3} and kcal mol^{-1} e^{-1}. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. The corresponding plots are shown in figure 5.4.

Potential & ID	Charge fit	q^{lr}	RMSE	MAE	MSE	m'	b'	R^2	c_V	c_q
Exact (Mult.)	All	All	3.75	2.42	-0.30	1.03	0.33	0.99	-0.1031	-19.415
/B|P|M|A		-1	8.01	6.42	-1.66	1.20	13.23	0.76		
		0	2.23	1.59	0.20	0.87	-0.40	0.76		
		1	3.96	3.32	-2.31	0.93	-7.01	0.82		
Exact (Mult.)	Individual	All	3.40	2.23	0.04	1.00	0.00	0.99		
/B|P|M|I		-1	7.10	5.76	-0.00	1.05	3.35	0.73	-0.1764	-25.985
		0	2.22	1.59	0.05	0.87	-0.53	0.76	-0.1039	1.000
		1	3.07	2.39	-0.00	0.83	-11.46	0.82	-0.1214	-14.143
Exact (NDDO)	All	All	4.00	2.90	0.40	0.93	-0.98	0.98	-0.1043	-15.389
/B|P|N|A		-1	7.32	6.38	3.15	0.86	-7.18	0.69		
		0	2.76	2.07	-0.60	0.74	-1.77	0.64		
		1	5.40	4.54	4.40	0.86	-5.06	0.82		
Exact (NDDO)	Individual	All	3.46	2.42	0.05	0.99	-0.14	0.99		
/B|P|N|I		-1	6.43	4.95	0.00	0.79	-15.29	0.69	-0.1390	-14.721
		0	2.66	1.96	0.06	0.72	-1.20	0.64	-0.1006	1.000
		1	3.10	2.49	-0.00	0.81	-12.66	0.82	-0.1133	-18.344
Point charge	All	All	3.70	2.65	0.06	1.02	0.51	0.99	-0.1028	-19.591
/B|P|P|A		-1	7.12	5.99	-0.96	1.23	16.26	0.81		
		0	2.70	2.02	0.42	0.79	-0.52	0.65		
		1	3.51	2.90	-1.34	0.90	-7.74	0.81		
Point charge	Individual	All	3.56	2.56	0.26	1.00	0.32	0.99		
/B|P|P|I		-1	6.69	5.68	-0.00	1.13	9.35	0.79	-0.1520	-23.939
		0	2.70	2.02	0.35	0.79	-0.58	0.65	-0.1032	1.000
		1	3.14	2.40	-0.00	0.82	-12.04	0.81	-0.1183	-15.728
Exact	All	All	2.04	1.43	-0.26	1.00	-0.35	1.00	-0.1025	-15.728
(*hr*-ref.)[24]		-1	3.07	2.46	0.01	1.10	7.18	0.94		
hr|B|E|A		0	1.56	1.13	-0.36	0.97	-0.47	0.89		
		1	2.98	2.10	0.02	0.96	-2.62	0.85		
Point charge	All	All	2.98	2.01	-0.56	1.04	0.42	0.99	-0.1009	-20.542
(*hr*-ref.)[24]		-1	5.27	3.92	-2.23	1.18	10.91	0.88		
hr|B|P|A		0	1.77	1.31	0.20	1.04	0.36	0.87		
		1	4.66	4.16	-3.61	0.94	-7.47	0.85		

RMSE compared to the PM6-EC-RISM model on the same set of geometries is small with a decrease of about $0.4\,\mathrm{kcal\,mol^{-1}}$ to an absolute value of $4.00\,\mathrm{kcal\,mol^{-1}}$. In contrast, the individual charge fit (/B|P|N|I) shows a more drastic improvement with an RMSE of $3.46\,\mathrm{kcal\,mol^{-1}}$, which is very similar to the corresponding multipole ESP model.

B3LYP optimised structures

Fitting these PMV correction models to the free energies obtained from the original B3LYP-PCM geometries leads to the results shown in figure 5.5 and table 5.5.

The use of B3LYP-PCM instead of PM6-PCM geometries in combination with the /B@PFL scheme leads to the same small increase in prediction quality as seen before for PM6-EC-RISM.

The multipole based model with one parameter set (/B|B|M|A) gives an RMSE of $3.58\,\mathrm{kcal\,mol^{-1}}$, while the individual charge fit (/B|B|M|I) gives $3.31\,\mathrm{kcal\,mol^{-1}}$. As before, the point charge models give similar but slightly worse results and also similar correction parameters to the multipole models. Again for this set of geometries, the NDDO-ESP with a single charge fit (/B|B|N|A) gives the worst results, especially for cations, although this can be corrected by fitting each charge state individually (/B|B|N|I).

As outlined above, the use of the B3LYP-PCM geometries allows a direct comparison of the /B@PFL models with the *hr*-reference. It is striking that although the *hr*-description of the solute improves the overall prediction quality, the models are still unable to accurately predict the solvation free energies for anionic species, as indicated by the high RMSE values in table 5.5. The reference models from Tielker et al.'s SAMPL6 publication also show larger deviations for ionic species, but the resulting RMSE and MAE are smaller. The difference between the /B@PFL models and the *hr* references is due to the use of the more approximate *lr*-ESP to calculate the solvent environment, as in both cases the solute is described by the *high* level theory, more specifically MP2.

When calculating solvation free energies, one effectively calculates the difference between the energy of the solute E_{sol} and its vacuum energy E_{vac}, which gives the effects of polarisation and conformational changes due to solvation. Adding the PMV-corrected excess chemical potential then gives an estimate of the solvation free energy. To accurately reproduce the experimental values with EC-RISM, it is therefore only necessary to accurately predict the difference between E_{sol} and E_{vac}, as systematic errors present in the absolute values may cancel out. In contrast, no such difference is computed for the excess chemical potential, and therefore only an indirect error cancellation through E_{sol} and E_{vac} is possible, and an accurate estimate of the ESP supplied to the 3D-RISM solver is required.

It can therefore be assumed that the use of the PM6-ESP within *lr*-EC-RISM or /B@PFL, whether in the form of the multipole, point charge or NDDO approximation, is not sufficient to provide accurate estimates of absolute solvation free energies of anionic

species.

It should be noted, however, that the main focus of this work is not on the prediction of solvation free energies, but rather on the prediction of pK_a values, where a cancellation of errors is possible not only between the energies of the solutes, but also between their excess chemical potentials. Therefore, in the following sections the effect of the parameterisation on the calculation of these quantities will be investigated.

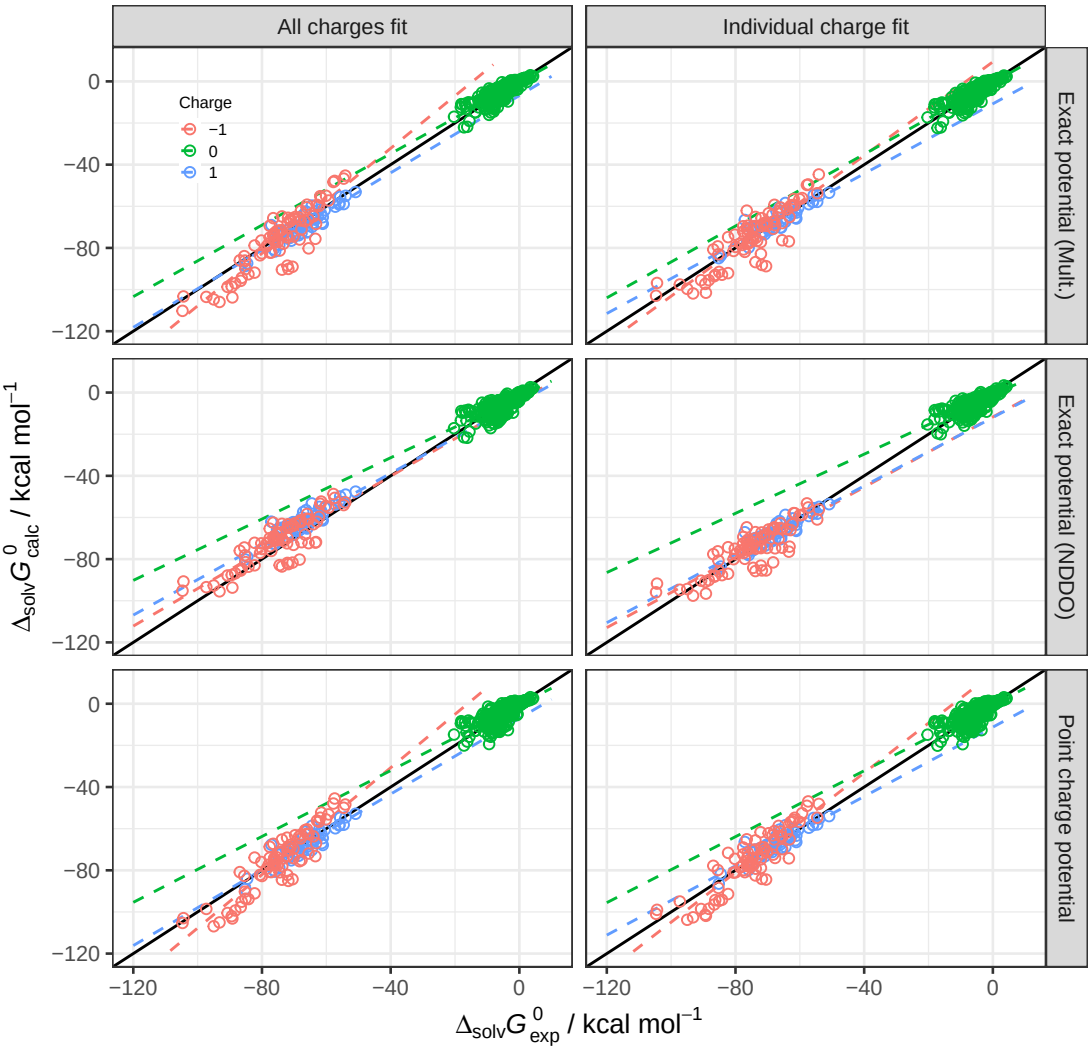

Figure 5.5.: Predicted versus experimental free energies of solvation for calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on the PCM conformers optimised at the B3LYP/6-311+G** level of theory. The first column shows predictions made with one set of PMV correction parameters for all charge states, while in the second column each state was parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Anionic, neutral and cationic species are shown in red, green and blue, respectively. The corresponding numerical model data are shown in table 5.5.

Table 5.5.: Statistical quantities and PMV correction parameters obtained for the free energy of solvation prediction on the MNSOL water subset. Single-point calculations were performed with ONIOM2(MP2/6-311+G**:PM6)-EC-RISM/B@PFL on B3LYP/6-311+G**-PCM-optimised structures. The statistical quantities RMSE, MAE and MSE have been calculated for the difference $\Delta\Delta_{\mathrm{solv}}G^0 = \Delta_{\mathrm{solv}}G^{0,/\mathrm{B@PFL}}_{\mathrm{calc}} - \Delta_{\mathrm{solv}}G^0_{\mathrm{exp}}$ and are given in kcal mol^{-1}. The parameters m' and b' were obtained by descriptive regression. The correction parameters c_V and c_q are given in kcal mol^{-1} Å^{-3} and kcal mol^{-1} e^{-1}. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. The corresponding plots are shown in figure 5.5.

Potential & ID	Charge fit	q^{lr}	RMSE	MAE	MSE	m'	b'	R^2	c_V	c_q
Exact (Mult.)	All	All	3.58	2.36	-0.21	1.03	0.37	0.99	-0.1024	-19.643
/B|B|M|A		-1	7.14	5.63	-1.25	1.26	18.26	0.81		
		0	2.17	1.57	0.24	0.86	-0.39	0.77		
		1	3.73	3.16	-1.97	0.93	-6.83	0.82		
Exact (Mult.)	Individual	All	3.31	2.21	0.08	1.00	0.10	0.99		
/B|B|M|I		-1	6.46	5.15	0.00	1.13	9.38	0.78	-0.1639	-25.157
		0	2.16	1.57	0.11	0.86	-0.50	0.77	-0.1031	1.000
		1	3.05	2.39	-0.00	0.84	-10.76	0.83	-0.1185	-15.073
Exact (NDDO)	All	All	4.06	2.98	0.54	0.93	-1.05	0.99	-0.1042	-14.947
/B|B|N|A		-1	6.81	5.90	3.32	0.90	-4.15	0.72		
		0	2.75	2.06	-0.67	0.74	-1.85	0.64		
		1	6.05	5.25	5.24	0.85	-4.73	0.83		
Exact (NDDO)	Individual	All	3.35	2.36	0.10	0.99	-0.04	0.99		
/B|B|N|I		-1	5.81	4.42	0.00	0.84	-11.71	0.72	-0.1327	-13.745
		0	2.60	1.91	0.14	0.71	-1.15	0.65	-0.0997	1.000
		1	3.02	2.43	0.00	0.82	-11.97	0.83	-0.1098	-19.292
Point charge	All	All	3.79	2.71	0.17	1.02	0.57	0.99	-0.1014	-19.966
/B|B|P|A		-1	6.71	5.52	-0.59	1.28	20.61	0.83		
		0	2.89	2.11	0.48	0.79	-0.46	0.62		
		1	3.30	2.75	-0.94	0.91	-6.99	0.82		
Point charge	Individual	All	3.70	2.64	0.33	1.01	0.45	0.99		
/B|B|P|I		-1	6.44	5.29	0.00	1.20	14.63	0.81	-0.1408	-23.363
		0	2.89	2.11	0.44	0.79	-0.49	0.62	-0.1016	1.000
		1	3.08	2.41	-0.00	0.83	-11.23	0.82	-0.1156	-16.714
Exact	All	All	2.04	1.43	-0.26	1.00	-0.35	1.00	-0.1025	-15.728
(*hr*-ref.)[24]		-1	3.07	2.46	0.01	1.10	7.18	0.94		
hr|B|E|A		0	1.56	1.13	-0.36	0.97	-0.47	0.89		
		1	2.98	2.10	0.02	0.96	-2.62	0.85		
Point charge	All	All	2.98	2.01	-0.56	1.04	0.42	0.99	-0.1009	-20.542
(*hr*-ref.)[24]		-1	5.27	3.92	-2.23	1.18	10.91	0.88		
hr|B|P|A		0	1.77	1.31	0.20	1.04	0.36	0.87		
		1	4.66	4.16	-3.61	0.94	-7.47	0.85		

5.2. Parameterisation of acidity constant corrections

5.2.1. Calculation of acidity constants

In order to accurately predict acidity constants, Tielker et al. used a second linear correction,[24] which must also be transferred to the ONIOM-EC-RISM context.

Given a deprotonation reaction

$$\mathrm{AH}_{n-i}^{-i} \rightleftharpoons \mathrm{AH}_{n-i-1}^{-i-1} + \mathrm{H}^{+} \tag{5.5}$$

of a molecule A with n titratable protons, the pK_a value, which is the main concern of this chapter and measures the acidity of a molecular system, is defined by

$$\mathrm{p}K_{\mathrm{a},i} = \frac{G_{i+1} + G_{\mathrm{H}^+} - G_i}{RT \ln 10} \tag{5.6}$$

with the macroscopic free energies of two adjacent protonation states i and $i+1$ of the solute and the free energy of the proton G_{H^+}.

In order to avoid the explicit calculation of the latter quantity in the original SAMPL6 publication, an empirical correction

$$\mathrm{p}K_{\mathrm{a,corr},i} = m\frac{G_{i+1} - G_i}{RT \ln 10} + b \tag{5.7}$$

was used with the parameters m and b, where the contributions from the proton are fitted into the intercept parameter b. The resulting model was fitted to experimental values from the Klicić pK_a data set in a similar way to equation 5.2.

Note that in the corresponding equation in the original SAMPL6 publication,[24] the free energy difference is reversed. However, this is an error and has been corrected in equation 5.7.

The macroscopic free energies used here can be obtained from the underlying microscopic states, i.e. the tautomers t, using a partition function approach

$$G_i = -RT \ln \sum_t \exp\left[-\beta G_{it}\right]. \tag{5.8}$$

Similarly, these microscopic free energies are obtained from the underlying conformers c via

$$G_{it} = -RT \ln \sum_c \exp\left[-\beta G_{\mathrm{sol},itc}^{\mathrm{corr}}\right]. \tag{5.9}$$

For the conformer free energies the PMV correction models discussed in the last section are applied. See ref. [122] for more details on this partition function approach.

The parameterisation of this model for the ONIOM-EC-RISM schemes leads to the same problems as the parameterisation of the corresponding PMV corrections. First,

Table 5.6.: Overview of all *lr*-PM6-EC-RISM pK_a correction models parameterised in this chapter. Table 5.7 is a continuation of this table and contains the corresponding /B@PFL-EC-RISM models. See section 5.2.3 for an explanation of the two pK_a fitting modes. Model IDs for the final pK_a models are obtained by adding the identifier for these two pK_a fitting models ("‖A" or "‖I") to the PMV correction IDs from table 5.1. Model IDs for thiol-free parameterisation, introduced later in this chapter, are identical to the naming scheme presented here, but include an additional "|nt" identifier (see tables 5.12 and 5.13). Model IDs are reported alongside the corresponding tables. The tables containing the model data are listed next to the model IDs in brackets.

EC-RISM	Optimisation	PMV correction model				pK_a correction model		
		Potential	Charge fit	Model ID	Tab.	pK_a fit	Model ID	Tab.
lr	PM6-PCM	Exact (Mult.)	All	*lr*|P|M|A	(5.2)	All	*lr*|P|M|A‖A	(5.8)
						Indiv.	*lr*|P|M|A‖I	(5.8)
			Indiv.	*lr*|P|M|I	(5.2)	All	*lr*|P|M|I‖A	(5.8)
						Indiv.	*lr*|P|M|I‖I	(5.8)
		Exact (NDDO)	All	*lr*|P|N|A	(5.2)	All	*lr*|P|N|A‖A	(5.8)
						Indiv.	*lr*|P|N|A‖I	(5.8)
			Indiv.	*lr*|P|N|I	(5.2)	All	*lr*|P|N|I‖A	(5.8)
						Indiv.	*lr*|P|N|I‖I	(5.8)
		Point charge	All	*lr*|P|P|A	(5.2)	All	*lr*|P|P|A‖A	(5.8)
						Indiv.	*lr*|P|P|A‖I	(5.8)
			Indiv.	*lr*|P|P|I	(5.2)	All	*lr*|P|P|I‖A	(5.8)
						Indiv.	*lr*|P|P|I‖I	(5.8)
	B3LYP-PCM	Exact (Mult.)	All	*lr*|B|M|A	(5.3)	All	*lr*|B|M|A‖A	(5.9)
						Indiv.	*lr*|B|M|A‖I	(5.9)
			Indiv.	*lr*|B|M|I	(5.3)	All	*lr*|B|M|I‖A	(5.9)
						Indiv.	*lr*|B|M|I‖I	(5.9)
		Exact (NDDO)	All	*lr*|B|N|A	(5.3)	All	*lr*|B|N|A‖A	(5.9)
						Indiv.	*lr*|B|N|A‖I	(5.9)
			Indiv.	*lr*|B|N|I	(5.3)	All	*lr*|B|N|I‖A	(5.9)
						Indiv.	*lr*|B|N|I‖I	(5.9)
		Point charge	All	*lr*|B|P|A	(5.3)	All	*lr*|B|P|A‖A	(5.9)
						Indiv.	*lr*|B|P|A‖I	(5.9)
			Indiv.	*lr*|B|P|I	(5.3)	All	*lr*|B|P|I‖A	(5.9)
						Indiv.	*lr*|B|P|I‖I	(5.9)

the size of the Klicić data set prevents a manual partitioning of its molecules. At the same time, this would introduce an arbitrary partitioning error as described above. To avoid this, the PFL can again be used to generate a correction model that does not require partitioning.

Since the size extrapolation limit of the /B scheme does not coincide with *hr*, it was argued earlier in this thesis that it would be useful to parameterise a new PMV correction, i.e. the /B@PFL parameterisation, in addition to the *hr* correction. See section 5.1 for a more detailed discussion. The same reasoning can be applied to the parameterisation of the pK_a correction. Therefore, in addition to the application of the *hr*-pK_a parameters, a new set of parameters will be parametrised at the /B@PFL level. As /B@PFL can be considered as a halfway point between *lr* and *hr*, an *lr*-pK_a correction will also be parameterised and tested in order to gain additional insight into the *high*-level description of the solute, as has been discussed previously for the PMV correction models.

For the same reasons as for the MNSOL calculations in section 5.1.1, the calculations are performed on the set of geometries from the Tielker et al. publication,[24] i.e. B3LYP-PCM-optimised structures as well as PM6-PCM re-optimised structures.

Tables 5.6 and 5.7 provide an overview of all pK_a correction models parameterised in this section and their underlying PMV corrections. The PMV model IDs introduced in table 5.1 are supplemented by two identifiers indicating the pK_a correction mode to give the pK_a model ID. In addition, an identifier can be added to indicate whether the parameterisation was performed on the full data set or on a subset excluding thiols. The reasons for this is presented in the following sections. "*hr*-MP2" refers to the parameters used in the SAMPL6 publication by Tielker et al.[24]

The pK_a fit columns in tables 5.6 and 5.7 indicate whether one set of pK_a correction parameters was used for the entire dataset ("All") or if acids and bases were corrected individually ("Indiv."). This results in two sets of pK_a correction parameters. Note that two types of individual correction therefore exist: The first is a PMV correction, where each charge state is parameterised individually, and the second is the aforementioned individual pK_a correction of acids and bases.

As already described for the PMV correction corresponding model parameters can be quickly retrieved by reffering to the tables given in tables 5.6 and 5.7. The model IDs introduced for the PMV correction are here augmented by an additional letter, that shows whether the "All" (A) or "Indiv." (I) corrections were applied.

5.2.2. Computational details

The optimisation and single point calculation strategy used for the pK_a calculations is similar to the previous calculations on the MNSOL dataset.[24] The settings used for EC-RISM and 3D-RISM calculations are identical to those used in the original SAMPL6 publication.[24]

Table 5.7.: Overview of all /B@PFL-EC-RISM pK_a correction models parameterised in this chapter. This table is a continuation of Table 5.6, which contains the corresponding *lr*-EC-RISM models and more details on the model ID naming scheme. Model IDs for thiol-free parameterisation, introduced later in this chapter, are identical to the naming scheme presented here, but include an additional "|nt" identifier (see tables 5.12 and 5.13). The tables containing the model data are listed next to the model IDs in brackets.

EC-RISM	Optimisation	PMV correction model				pK_a correction model		
		Potential	Charge fit	Model ID	Tab.	pK_a fit	Model ID	Tab.
/B@PFL	PM6-PCM	Exact (Mult.)	All	/B|P|M|A	(5.4)	All	/B|P|M|A‖A	(5.10)
						Indiv.	/B|P|M|A‖I	(5.10)
			Indiv.	/B|P|M|I	(5.4)	All	/B|P|M|I‖A	(5.10)
						Indiv.	/B|P|M|I‖I	(5.10)
		Exact (NDDO)	All	/B|P|N|A	(5.4)	All	/B|P|N|A‖A	(5.10)
						Indiv.	/B|P|N|A‖I	(5.10)
			Indiv.	/B|P|N|I	(5.4)	All	/B|P|N|I‖A	(5.10)
						Indiv.	/B|P|N|I‖I	(5.10)
		Point charge	All	/B|P|P|A	(5.4)	All	/B|P|P|A‖A	(5.10)
						Indiv.	/B|P|P|A‖I	(5.10)
			Indiv.	/B|P|P|I	(5.4)	All	/B|P|P|I‖A	(5.10)
						Indiv.	/B|P|P|I‖I	(5.10)
	B3LYP-PCM	Exact (Mult.)	All	/B|B|M|A	(5.5)	All	/B|B|M|A‖A	(5.11)
						Indiv.	/B|B|M|A‖I	(5.11)
			Indiv.	/B|B|M|I	(5.5)	All	/B|B|M|I‖A	(5.11)
						Indiv.	/B|B|M|I‖I	(5.11)
		Exact (NDDO)	All	/B|B|N|A	(5.5)	All	/B|B|N|A‖A	(5.11)
						Indiv.	/B|B|N|A‖I	(5.11)
			Indiv.	/B|B|N|I	(5.5)	All	/B|B|N|I‖A	(5.11)
						Indiv.	/B|B|N|I‖I	(5.11)
		Point charge	All	/B|B|P|A	(5.5)	All	/B|B|P|A‖A	(5.11)
						Indiv.	/B|B|P|A‖I	(5.11)
			Indiv.	/B|B|P|I	(5.5)	All	/B|B|P|I‖A	(5.11)
						Indiv.	/B|B|P|I‖I	(5.11)

The B3LYP-PCM optimised structures from the *hr*-reference publication [24] were reoptimised using Gaussian16 with the PM6-PCM level. Additionally, the optimisation strategy as described in section 5.1.2 was applied.

Single point calculations were performed at the PM6-EC-RISM and ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL levels of theory. For each level of theory three ESP variants were used, i.e. the multipole, NDDO and point charge based ESP variants were fed to the 3D-RISM solver to calculate the solvent structure.

As also stated in section 5.1.2, the PM6-EC-RISM calculations were performed with looser EMPIRE convergence criteria than the corresponding ONIOM-EC-RISM calculations. It has been validated that this does not lead to large differences in the solvent structure, as estimated by comparison of the resulting excess chemical potential. The corresponding results are given in table 2 and figure 2 in the appendix.

To ensure identical parameterisation with respect to previous models, the Mathematica notebooks used for the SAMPL6 publication were provided by Nicolas Tielker and only modified to allow the use of the ONIOM-EC-RISM input. The associated raw data, such as atomic coordinate files, can be found in the electronic supplementary material.[116]

5.2.3. Acidity constant prediction with PM6-EC-RISM

PM6 reoptimised structures

The pK_a prediction results and model parameters m and b for EC-RISM at the PM6 level of theory on the reoptimised PM6-PCM geometries are shown in figure 5.6 and table 5.8.

As before, the statistical quantities RMSE, MAE and MSE were calculated for the difference $\Delta\mathrm{p}K_\mathrm{a} = \mathrm{p}K_\mathrm{a,calc} - \mathrm{p}K_\mathrm{a,exp}$. In addition, descriptive linear regression was performed, yielding m' and b'. Note that these parameters were obtained using the equation $\mathrm{p}K_\mathrm{a,corr} = m'\mathrm{p}K_\mathrm{a,exp} + b'$, as this is the procedure employed in the reference publication by Tielker et al.[24] for the SAMPL6 dataset. However, it is more common to use the equation $\mathrm{p}K_\mathrm{a,exp} = m''\mathrm{p}K_\mathrm{a,corr} + b''$, which, in this case, yields $m'' = 1$ and $b'' = 0$ due to its similarity with the pK_a correction equation (equation 5.7). Nevertheless, the definition used here can still be employed alongside the other prediction measures to assess the model's performance, since an ideal model would yield $m' = 1$ and $b' = 0$, and any deviations would indicate that the model is less than ideal.

To allow a comparison between these pK_a models and the results from the *hr* level of theory, the best performing models from the Tielker et al. publication are also shown in the table. In the case of using the exact potential, two models were reported by the authors. A pre-submission model where some non-converged single point calculations were replaced by the corresponding point charge based calculations and a second post-submission model where all calculations could be converged using the exact potential.

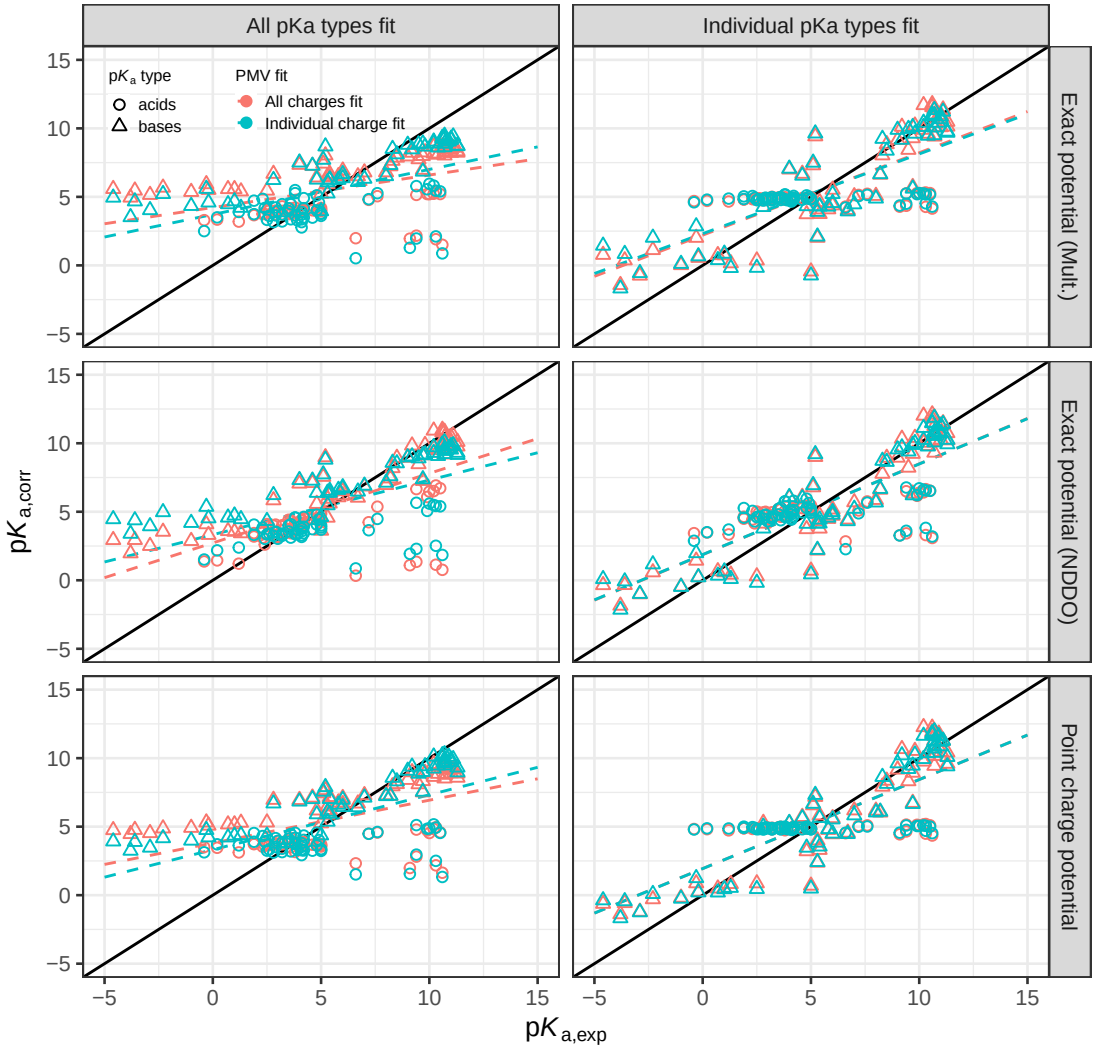

Figure 5.6.: Predicted versus experimental pK_a values for calculations at the PM6-EC-RISM level of theory on the reoptimised PM6-PCM conformers. The first column shows predictions made with one set of pK_a correction parameters for both pK_a types, i.e. acids and bases, while in the second column each type has been parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). table 5.8 shows the corresponding numerical model data.

Although the first model was reported with a slightly lower RMSE for the Klicić dataset, subsequent validation on the independent SAMPL6 dataset showed that the second model was superior.[24] Therefore, this second model was chosen as the reference model for the following discussions.

As the free energies obtained from the single point calculations are PMV corrected using the previously generated models, pK_a models were parameterised for both of these correction schemes. As these calculations were performed on PM6-PCM geometries, the PMV correction models parameterised for the same level of optimisation were applied.

The resulting multipole and point-charged-based models, using a single set of parameters for all charge states (lr|P|M|A‖A), show an RMSE of 3.40 and 3.23 respectively. These similar results are consistent with the previous solvation free energy predictions where these ESP types consistently gave results similar to those observed here. The corresponding NDDO model (lr|P|N|A‖A) reduces the RMSE to 2.73. Applying individual corrections for each charge state results in a slight decrease in RMSE and MAE for all except the NDDO model.

However, the data plotted in figure 5.6 and the descriptive regression show that all these initial models are unable to accurately predict pK_a values for acids. An ideal model would give a regression parameter m' of 1, while the corresponding intercept must be zero. Instead, the slope parameter of the descriptive regression for acids is close to zero. This effect is particularly pronounced for multipole and point charge based models, with a m' of about 0.03 and a low R^2 of about 0.01 in both cases. As a consequence, the RMSE gives little or no information about the model performance.

In the previous sections it was shown that the PM6-EC-RISM models are unable to accurately predict solvation free energies for anions. Here, "bases" refers to compounds that undergo deprotonation from the charge states +1 to 0, while "acids" are defined in the same way as compounds that change charge states from 0 to -1. It is therefore reasonable to assume that the poor prediction quality for anions at this level of theory is the cause of the equally poor prediction quality for acids.

In an attempt to circumvent this problem, additional pK_a models were generated where each pK_a type, i.e. acids and bases, is parameterised with individual sets of correction parameters m and b. The resulting parameters and statistical quantities are also shown in table 5.8. The corresponding pK_a values are shown in figure 5.6.

This approach leads in all cases to a drastic improvement in prediction quality for bases compared to the corresponding model with the same PMV correction but a single set of pK_a correction parameters. For example, the m' parameter and the RMSE can be improved from 0.56 to 0.86 and from 2.49 to 1.72 respectively for the NDDO model using a single set of PMV ocrrection parameters (lr|P|N|A‖I).

However, the descriptive regression parameters show that it is still not possible to obtain meaningful results for acids. Each of these additional models shows an R^2 below 0.12 for these compounds. The worst models are obtained for the multipole and point charge based ESPs, where the regression parameter m for acids is close to zero, giving

Table 5.8.: Statistical quantities and pK_a correction parameters for PM6-EC-RISM on PM6-PCM reoptimised geometries from the Klicić data set. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. Omitted values were not reported. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 5.6 for the corresponding plots. Refer to table 5.2 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	3.40	2.43	0.00	0.24	4.22	0.24	0.20	18.86
		lr|P|M|A‖A	acids	3.06	2.01	-0.97	0.03	3.78	0.01		
			bases	3.71	2.84	0.94	0.23	5.70	0.81		
	All	Individual	All	2.46	1.88	-0.00	0.60	2.21	0.60		
		lr|P|M|A‖I	acids	2.84	2.23	0.00	0.01	4.84	0.01	0.06	9.43
			bases	2.03	1.55	-0.00	0.81	1.20	0.81	0.70	47.63
	Individual	All	All	3.19	2.27	-0.00	0.33	3.71	0.33	0.31	27.08
		lr|P|M|I‖A	acids	3.15	2.17	-0.97	0.03	3.74	0.01		
			bases	3.23	2.36	0.94	0.35	4.92	0.78		
	Individual	Individual	All	2.52	1.90	-0.00	0.58	2.32	0.58		
		lr|P|M|I‖I	acids	2.84	2.24	0.00	0.01	4.85	0.01	0.06	9.37
			bases	2.17	1.58	-0.00	0.78	1.37	0.78	0.69	49.96
Exact (NDDO)	All	All	All	2.73	1.70	-0.00	0.51	2.73	0.51	0.47	37.60
		lr|P|N|A‖A	acids	2.97	1.70	-1.02	0.16	3.06	0.09		
			bases	2.49	1.69	0.98	0.56	3.72	0.86		
	All	Individual	All	2.26	1.74	0.00	0.66	1.86	0.66		
		lr|P|N|A‖I	acids	2.71	2.14	0.00	0.09	4.42	0.09	0.27	24.42
			bases	1.72	1.36	-0.00	0.86	0.86	0.86	0.72	53.32
	Individual	All	All	3.02	2.03	-0.00	0.40	3.33	0.40	0.34	29.21
		lr|P|N|I‖A	acids	2.96	1.91	-1.24	0.12	3.06	0.11		
			bases	3.08	2.14	1.19	0.41	4.86	0.85		
	Individual	Individual	All	2.27	1.71	-0.00	0.66	1.88	0.66		
		lr|P|N|I‖I	acids	2.69	2.12	-0.00	0.11	4.33	0.11	0.32	28.90
			bases	1.77	1.31	-0.00	0.85	0.92	0.85	0.72	52.00
Point charge	All	All	All	3.23	2.27	0.00	0.31	3.81	0.31	0.21	19.29
		lr|P|P|A‖A	acids	3.11	2.09	-1.17	0.01	3.65	0.00		
			bases	3.34	2.44	1.13	0.33	5.27	0.88		
	All	Individual	All	2.31	1.77	0.00	0.65	1.94	0.65		
		lr|P|P|A‖I	acids	2.84	2.24	0.00	0.00	4.86	0.00	0.05	8.82
			bases	1.63	1.32	0.00	0.88	0.77	0.88	0.55	38.21
	Individual	All	All	3.02	2.07	-0.00	0.40	3.32	0.40	0.29	25.01
		lr|P|P|I‖A	acids	3.16	2.18	-1.17	0.01	3.64	0.00		
			bases	2.87	1.96	1.13	0.45	4.51	0.88		
	Individual	Individual	All	2.30	1.74	0.00	0.65	1.94	0.65		
		lr|P|P|I‖I	acids	2.85	2.24	0.00	0.00	4.87	0.00	0.04	8.03
			bases	1.62	1.26	-0.00	0.88	0.77	0.88	0.56	40.49
Exact (*hr*-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A‖A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (*hr*-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A‖A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

a model that almost gives a constant pK_a value prediction for any input of free energies calculated by the solvation model.

B3LYP optimised structures

To rule out the possibility that these poor results are due to the use of the PM6 conformers, additional single point calculations and pK_a model parameterisations were performed on the original set of B3LYP-PCM-optimised geometries. The results are shown in figure 5.7 and table 5.9. In contrast to the previous models, the PMV correction parameters obtained for the B3LYP-PCM optimisation level were used here.

However, it can be seen that this set of geometries leads to only small improvements. From figure 5.7 alone it is clear that changing the level of optimisation does not have a significant effect and does not solve the underlying problem of poor description of anionic species.

In conclusion, the description of both the solute and the solvent environment at the PM6 level of theory cannot be used to obtain a meaningful pK_a prediction for acids, as defined here through the change from the charge state 0 to -1. As the SAMPL6 data set, which is used as a test data set for these models, contains a wide range of charge states, including negatively charged states, these models are discarded.

5.2.4. Acidity constant prediction with ONIOM-EC-RISM/B@PFL

PM6 reoptimised structures

The consequence of changing the solvation model from PM6-EC-RISM to ONIOM-EC-RISM/B@PFL is that the solute is modelled at the *high* level of theory, i.e. in this case MP2, while the solvent environment is still modelled at the less expensive *low* level of theory. The results of these ONIOM-EC-RISM calculations on the reoptimised PM6-PCM geometries are shown in table 5.10 and figure 5.8. For the sake of consistency, the PMV-correction parameters from the PM6-PCM reoptimised structures were again used for this set of geometries.

For the multipole-based ESP using the classic correction scheme from the SAMPL6 publications, i.e. one set of parameters for all charge states in the PMV-correction and one set of parameters correction for all pK_a-types in the pK_a correction (/B|P|M|A‖A), results in an RMSE of 2.07 and an m' of 0.72. Hence, this correction model shows pronounced deviations from the ideal model but a significantly enhanced R^2 as compared to the previous PM6-EC-RISM model. Bases demonstrated considerably superior results with a value of 1.58 when contrasted with acids, which achieved a value of 2.47. The final outcome is therefore significantly influenced by the prediction quality of the acids. It is noteworthy that the mean squared error for acids is -0.63, predicting values that are too small, whereas the MSE for bases is +0.60, predicting values that are too large.

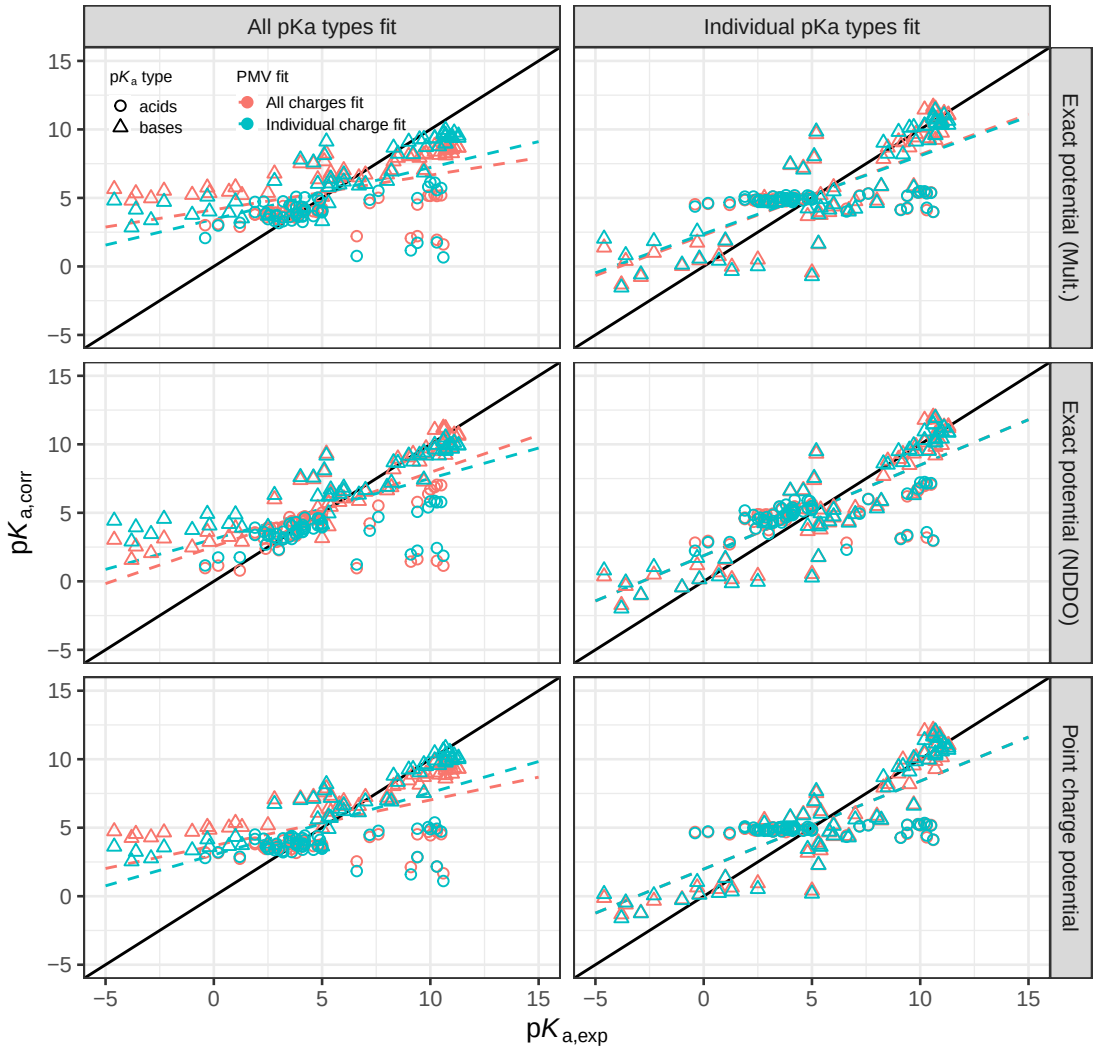

Figure 5.7.: Predicted versus experimental pK_a values for calculations at the PM6-EC-RISM level of theory on the PCM conformers optimised at the B3LYP/6-311+G** level of theory. The first column shows predictions made with one set of pK_a correction parameters for both pK_a types, i.e. acids and bases, while in the second column each type was parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). Table 5.9 shows the corresponding numerical model data.

Table 5.9.: Statistical quantities and pK_a correction parameters for PM6-EC-RISM on the B3LYP-PCM optimised geometries from the Klicić data set. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. Omitted values were not reported. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 5.7 for the corresponding plots. Refer to table 5.3 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	3.37	2.38	-0.00	0.25	4.14	0.25	0.21	19.40
		lr|B|M|A||A	acids	3.05	1.96	-1.01	0.03	3.71	0.01		
			bases	3.65	2.79	0.98	0.25	5.63	0.79		
	All	Individual	All	2.50	1.89	-0.00	0.59	2.28	0.59		
		lr|B|M|A||I	acids	2.83	2.23	0.00	0.01	4.82	0.01	0.08	11.08
			bases	2.13	1.57	-0.00	0.79	1.32	0.79	0.66	45.36
	Individual	All	All	3.07	2.11	-0.00	0.38	3.45	0.38	0.35	30.02
		lr|B|M|I||A	acids	3.12	2.02	-0.96	0.05	3.67	0.01		
			bases	3.03	2.19	0.92	0.42	4.52	0.76		
	Individual	Individual	All	2.55	1.94	-0.00	0.57	2.38	0.57		
		lr|B|M|I||I	acids	2.83	2.23	-0.00	0.01	4.81	0.01	0.10	12.05
			bases	2.26	1.67	-0.00	0.76	1.49	0.76	0.64	47.80
Exact (NDDO)	All	All	All	2.64	1.66	-0.00	0.54	2.54	0.54	0.48	38.80
		lr|B|N|A||A	acids	2.85	1.67	-0.98	0.21	2.89	0.14		
			bases	2.42	1.64	0.94	0.59	3.47	0.84		
	All	Individual	All	2.26	1.72	-0.00	0.66	1.87	0.66		
		lr|B|N|A||I	acids	2.64	2.07	-0.00	0.14	4.17	0.14	0.33	29.31
			bases	1.84	1.38	-0.00	0.84	0.99	0.84	0.68	51.32
	Individual	All	All	2.91	1.90	0.00	0.44	3.08	0.44	0.37	31.25
		lr|B|N|I||A	acids	2.88	1.78	-1.24	0.16	2.86	0.17		
			bases	2.94	2.01	1.19	0.45	4.56	0.83		
	Individual	Individual	All	2.27	1.71	0.00	0.66	1.87	0.66		
		lr|B|N|I||I	acids	2.60	2.03	-0.00	0.17	4.05	0.17	0.39	34.30
			bases	1.90	1.39	0.00	0.83	1.05	0.83	0.68	49.92
Point charge	All	All	All	3.18	2.20	-0.00	0.33	3.69	0.33	0.22	19.99
		lr|B|P|A||A	acids	3.10	2.05	-1.19	0.02	3.59	0.01		
			bases	3.25	2.35	1.15	0.35	5.14	0.86		
	All	Individual	All	2.33	1.77	-0.00	0.64	1.98	0.64		
		lr|B|P|A||I	acids	2.84	2.23	-0.00	0.01	4.84	0.01	0.08	11.12
			bases	1.70	1.33	0.00	0.86	0.84	0.86	0.53	37.13
	Individual	All	All	2.88	1.91	0.00	0.45	3.02	0.45	0.32	27.50
		lr|B|P|I||A	acids	3.12	2.07	-1.14	0.02	3.62	0.01		
			bases	2.63	1.75	1.10	0.52	4.03	0.86		
	Individual	Individual	All	2.33	1.76	-0.00	0.64	1.98	0.64		
		lr|B|P|I||I	acids	2.84	2.24	-0.00	0.01	4.84	0.01	0.09	11.62
			bases	1.70	1.30	-0.00	0.86	0.84	0.86	0.53	39.51
Exact (*hr*-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A||A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (*hr*-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A||A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

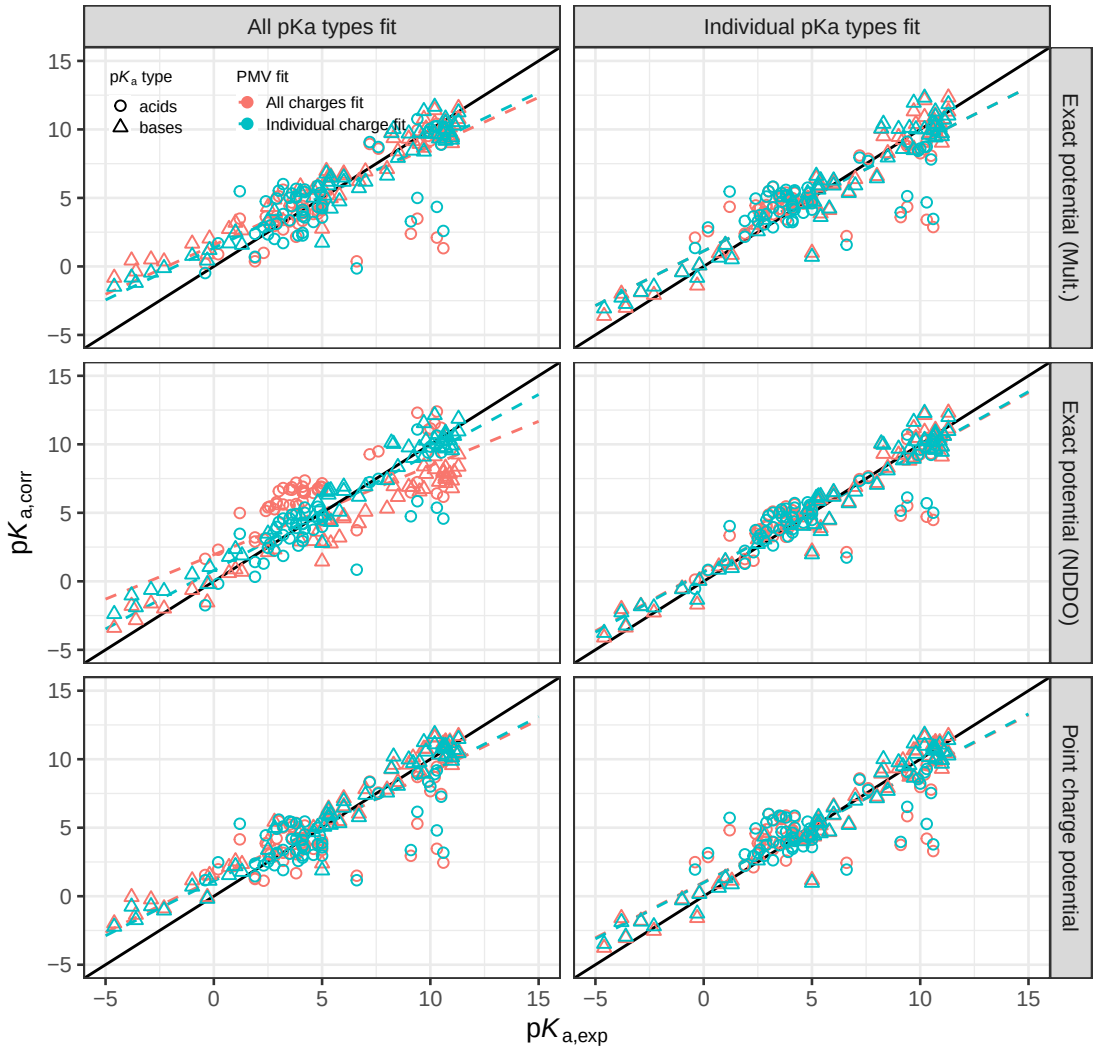

Figure 5.8.: Predicted versus experimental pK_a values for calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on the reoptimised PM6-PCM conformers. The first column shows predictions made with one set of pK_a correction parameters for both pK_a types, i.e. acids and bases, while in the second column each type has been parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). Table 5.10 shows the corresponding numerical model data.

This tendency is consistent throughout all models with one set of pK_a parameters for all pK_a types.

This behaviour, and in particular the very different prediction quality for acids and bases, justifies the application of two separate corrections. By applying this pK_a correction approach to the previously discussed model (/B|P|M|A‖I), the RMSE can be improved to a value of 1.75. The clear difference between the correction parameters m and b for acids and bases is striking. The parameters for bases are very similar to the *hr*-references for the exact potential and, as expected, the values for acids are significantly different, due to the larger deviations observed for acids before. The pK_a correction is a linear function of $\Delta G_i = G_{i+1} - G_i$. Since fitting the model to the experimental data of the Klicić dataset produces similar parameters to the reference, it can be concluded that the /B@PFL approach for bases is able to extrapolate to the ΔG_i of the *hr*-reference. Conversely, the ΔG_i for acids cannot be reproduced, resulting in significantly different parameters.

The model with a single pK_a correction and an individual PMV correction (/B|P|M|I‖A) gives an RMSE of 1.92, which is a slight improvement over the corresponding model with a single PMV correction. However, the same problems as before remain for acids. The application of the individual pK_a correction (/B|P|M|I‖I) here gives an RMSE of 1.76 and thus a similar improvement to the previously discussed model with a single PMV correction. The correction parameters are also very similar to this model.

A similar picture emerges for the NDDO-based potential. The best of these models with an RMSE of 1.36 results from the application of individual corrections for both the PMV and the pK_a correction (/B|P|N|I‖I) and represents an improvement over the corresponding model with a single pK_a correction. The second best results are obtained with a single PMV correction and the individual pK_a correction (/B|P|N|A‖I) with an RMSE of 1.41. The significantly poorer prediction quality of the acids is also decisive for the overall prediction quality of the NDDO models.

It is also noticeable that the NDDO parameterisations of the bases lead to an almost identical slope parameter m, but to a significantly different b compared to the *hr*-reference model and the corresponding ONIOM models with the multipole based potential. It can therefore be assumed that the ONIOM models that employ the NDDO-potential extrapolate to a different ΔG_i than those models.

Although the point charge models for the PMV parameterisation on the MNSOL data set always yielded similar results to the calculations with the multipole-based ESP, a significant difference can be seen for the pK_a calculations discussed here. Slightly smaller RMSEs are obtained for all point charge models compared to the corresponding multipole calculations. Again, the model with two individual corrections (/B|P|P|I‖I) gives the best result. It is noteworthy that this correction approach gives slightly better predictions than the reference point-charge model from the original SAMPL6 publication. The base correction parameters resulting from the single pK_a correction models (/B|P|N|A‖A and /B|P|N|I‖A) are very similar to this reference model. Therefore, for

Table 5.10.: Statistical quantities and pK_a correction parameters for ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL on the PM6-PCM reoptimised geometries from the Klicić data set. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. Omitted values were not reported. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 5.8 for the corresponding plots. Refer to table 5.4 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	2.07	1.34	-0.00	0.72	1.56	0.72	0.57	-113.07
		/B|P|M|A‖A	acids	2.47	1.43	-0.63	0.54	1.60	0.37		
			bases	1.58	1.25	0.60	0.74	2.24	0.95		
	All	Individual	All	1.75	1.16	-0.00	0.80	1.12	0.80		
		/B|P|M|A‖I	acids	2.26	1.52	-0.00	0.37	3.06	0.37	0.39	-75.65
			bases	1.07	0.81	-0.00	0.95	0.33	0.95	0.73	-148.02
	Individual	All	All	1.92	1.36	-0.00	0.76	1.34	0.76	0.62	-123.90
		/B|P|M|I‖A	acids	2.36	1.58	-0.24	0.58	1.79	0.40		
			bases	1.36	1.14	0.23	0.80	1.48	0.93		
	Individual	Individual	All	1.76	1.24	0.00	0.80	1.12	0.80		
		/B|P|M|I‖I	acids	2.20	1.58	0.00	0.40	2.91	0.40	0.43	-83.77
			bases	1.18	0.91	0.00	0.93	0.40	0.93	0.73	-146.45
Exact (NDDO)	All	All	All	2.31	2.09	0.00	0.65	1.94	0.65	0.56	-110.82
		/B|P|N|A‖A	acids	2.46	2.32	1.71	0.64	3.47	0.61		
			bases	2.15	1.87	-1.64	0.74	-0.04	0.96		
	All	Individual	All	1.41	0.94	-0.00	0.87	0.72	0.87		
		/B|P|N|A‖I	acids	1.77	1.15	-0.00	0.61	1.88	0.61	0.54	-108.02
			bases	0.93	0.73	-0.00	0.96	0.25	0.96	0.73	-143.40
	Individual	All	All	1.48	1.01	0.00	0.86	0.80	0.86	0.66	-131.07
		/B|P|N|I‖A	acids	1.76	1.10	-0.44	0.75	0.79	0.65		
			bases	1.14	0.92	0.43	0.86	1.26	0.96		
	Individual	Individual	All	1.36	0.93	-0.00	0.88	0.68	0.88		
		/B|P|N|I‖I	acids	1.67	1.11	-0.00	0.65	1.68	0.65	0.58	-113.95
			bases	0.97	0.75	0.00	0.96	0.27	0.96	0.73	-146.03
Point charge	All	All	All	1.85	1.27	0.00	0.78	1.24	0.78	0.54	-107.36
		/B|P|P|A‖A	acids	2.29	1.55	-0.66	0.46	1.97	0.41		
			bases	1.28	1.01	0.64	0.84	1.60	0.95		
	All	Individual	All	1.68	1.13	-0.00	0.81	1.03	0.81		
		/B|P|P|A‖I	acids	2.18	1.52	0.00	0.41	2.87	0.41	0.48	-95.11
			bases	0.98	0.76	-0.00	0.95	0.28	0.95	0.61	-122.92
	Individual	All	All	1.75	1.19	-0.00	0.80	1.11	0.80	0.57	-112.67
		/B|P|P|I‖A	acids	2.21	1.54	-0.51	0.48	2.05	0.44		
			bases	1.13	0.87	0.49	0.88	1.22	0.96		
	Individual	Individual	All	1.65	1.10	0.00	0.82	0.99	0.82		
		/B|P|P|I‖I	acids	2.14	1.53	-0.00	0.44	2.75	0.44	0.52	-102.36
			bases	0.96	0.67	0.00	0.96	0.27	0.96	0.61	-123.35
Exact (*hr*-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A‖A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (*hr*-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A‖A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

the same reasons as before, it can be assumed that they can be extrapolated to the same ΔG_i as the reference model.

Figure 5.8 shows that there is a group of molecules which systematically give worse results than the rest of the Klicić data set. These are at experimental values of about 10 pK_a units, but values of about 2.5 to 7.5 are predicted here. These molecules are assigned to the class of "thiols" in the Klicić data set. The previously used classification of the pK_a-type of acids based on the transition between the charge states 0 to -1 covers the substance classes "acids", "phenols" and "thiols" of the Klicić data set. When the RMSE, MAE and MSE are calculated separately for these and all other classes of compounds, the largest RMSEs and MAEs are obtained for thiols for all models. These values are shown in tables 3, 4 and 5 in the appendix. The MSE also indicates that the thiol pK_a values are underestimated by all models. In contrast, the other two classes of acids, "phenols" and "acids", result in significantly smaller deviations from the experimental reference, so that it can be assumed that the lack of description of the thiols is primarily responsible for the poor results of the acids and thus also of the models with a single pK_a correction.

B3LYP optimised structures

These findings also apply to the results obtained from the B3LYP-PCM-optimised structures shown in figure 5.9 and table 5.11.

Again, the class of thiols can be identified as the cause of the significantly poorer predictions of the models, as shown in tables 6, 7 and 8 in the appendix. The best overall result for these geometries is again obtained by applying individual fits for both the PMV correction and the pK_a correction to the NDDO single point energies (/B|B|N|I‖I). This gives an RMSE of 1.56, while the corresponding model with a single PMV correction (/B|B|N|A‖I) gives a slightly higher value of 1.61.

The correction parameters obtained from these geometries are also very similar to the previous models. It should be noted, however, that they give slightly worse pK_a predictions compared to the PM6-PCM reoptimised structures, with an increase in RMSE of about 0.2 pK_a units. Although these results were worse than the previous geometries, the original results by Tielker et al. also showed worse results for the expected better post-submission model. The final evaluation of the predictive power is therefore carried out in the following sections using the independent test data set from the SAMPL6 challenge.

Before proceeding, it is worth summarising the main results. All models show significantly worse results for acids, defined as the transition between charge states 0 and -1, compared to bases. The prediction of solvation free energies for the MNSOL dataset and the definition of acids based on charge states suggest that the poor predictive power for anions is the main reason for the poor results for acids. Calculation of the statistical measures by substance class shows that these deviations can also be attributed to the

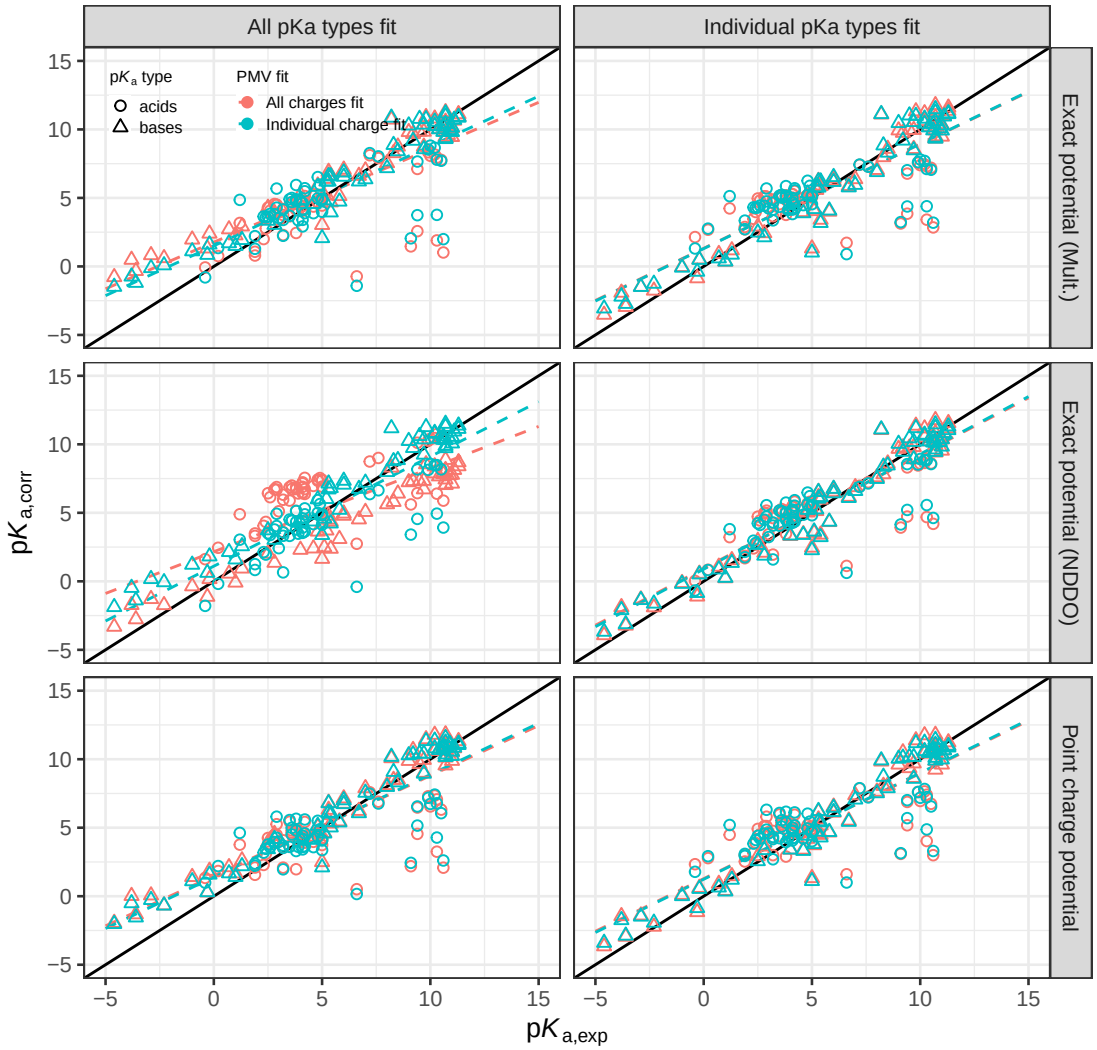

Figure 5.9.: Predicted versus experimental pK_a values for calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on the PCM conformers optimised at the B3LYP/6-311+G** level of theory. The first column shows predictions made with one set of pK_a correction parameters for both pK_a types, i.e. acids and bases, while in the second column each type has been parameterised separately. The rows represent the ESP supplied to the 3D-RISM solver, as further explained in the main text. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). Table 5.11 shows the corresponding numerical model data.

Table 5.11.: Statistical quantities and pK_a correction parameters for ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL on the B3LYP-PCM optimised geometries from the Klicić data set. The best performing *hr*-reference models from the original SAMPL6 publication[24] are shown in the last rows. Omitted values were not reported. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 5.9 for the corresponding plots. Refer to table 5.5 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	2.20	1.36	0.00	0.68	1.77	0.68	0.58	-117.48
		/B|B|M|A‖A	acids	2.71	1.54	-0.71	0.36	2.43	0.23		
			bases	1.57	1.18	0.68	0.75	2.25	0.95		
	All	Individual	All	1.90	1.31	-0.00	0.76	1.32	0.76		
		/B|B|M|A‖I	acids	2.50	1.85	0.00	0.23	3.76	0.23	0.38	-73.47
			bases	1.05	0.78	-0.00	0.95	0.32	0.95	0.74	-151.84
	Individual	All	All	2.03	1.35	0.00	0.73	1.51	0.73	0.65	-130.13
		/B|B|M|I‖A	acids	2.58	1.67	-0.40	0.39	2.59	0.26		
			bases	1.31	1.04	0.38	0.82	1.50	0.94		
	Individual	Individual	All	1.89	1.32	0.00	0.76	1.30	0.76		
		/B|B|M|I‖I	acids	2.45	1.81	0.00	0.26	3.60	0.26	0.44	-86.00
			bases	1.11	0.85	-0.00	0.94	0.36	0.94	0.74	-150.91
Exact (NDDO)	All	All	All	2.43	2.18	-0.00	0.61	2.16	0.61	0.57	-113.21
		/B|B|N|A‖A	acids	2.68	2.44	1.71	0.49	4.21	0.48		
			bases	2.17	1.93	-1.65	0.74	-0.05	0.95		
	All	Individual	All	1.61	1.12	0.00	0.83	0.94	0.83		
		/B|B|N|A‖I	acids	2.06	1.48	0.00	0.48	2.56	0.48	0.55	-111.87
			bases	0.98	0.76	-0.00	0.95	0.28	0.95	0.73	-145.62
	Individual	All	All	1.74	1.17	0.00	0.80	1.11	0.80	0.65	-130.25
		/B|B|N|I‖A	acids	2.11	1.35	-0.68	0.55	1.49	0.51		
			bases	1.29	1.00	0.66	0.84	1.62	0.95		
	Individual	Individual	All	1.56	1.09	-0.00	0.84	0.89	0.84		
		/B|B|N|I‖I	acids	1.99	1.42	-0.00	0.51	2.37	0.51	0.60	-119.63
			bases	0.99	0.78	-0.00	0.95	0.28	0.95	0.73	-148.93
Point charge	All	All	All	2.02	1.33	-0.00	0.73	1.49	0.73	0.55	-110.23
		/B|B|P|A‖A	acids	2.57	1.71	-0.67	0.28	2.86	0.24		
			bases	1.29	0.96	0.64	0.85	1.57	0.95		
	All	Individual	All	1.88	1.30	0.00	0.77	1.28	0.77		
		/B|B|P|A‖I	acids	2.48	1.86	0.00	0.24	3.68	0.24	0.49	-96.40
			bases	1.00	0.76	-0.00	0.95	0.29	0.95	0.62	-125.16
	Individual	All	All	1.95	1.31	-0.00	0.75	1.39	0.75	0.57	-114.32
		/B|B|P|I‖A	acids	2.52	1.73	-0.60	0.28	2.90	0.27		
			bases	1.18	0.91	0.57	0.88	1.33	0.96		
	Individual	Individual	All	1.84	1.27	-0.00	0.78	1.24	0.78		
		/B|B|P|I‖I	acids	2.44	1.84	-0.00	0.27	3.58	0.27	0.54	-106.67
			bases	0.96	0.71	-0.00	0.96	0.27	0.96	0.62	-125.94
Exact (*hr*-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A‖A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (*hr*-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A‖A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

substance class "thiols". Separate parameterisation of acids and bases leads to a significant improvement of the results. The resulting parameters obtained from multipole-based potential for bases are very similar to those of the exact potential *hr*-reference model. Similarly, the point charge model provides similar parameter models to the *hr* point charge reference. Only the NDDO potential based model gives different model parameters with larger deviations from the reference model. Since the pK_a correction is only a function of the difference ΔG_i, it can be concluded that the multipole and point charge models extrapolate to the ΔG_i of the respective reference. In contrast, the NDDO model extrapolates to a different solute free energy difference.

Evaluation of thiol-free models

As the substance class "thiols" was identified as a clear outlier that strongly influences the predictive power for acids, it can also be assumed that any remaining errors are masked. In order to further assess the predictive power of the ONIOM-EC-RISM/B@PFL level of theory, additional models were created and parameterised on the Klicić dataset without the substance class "thiols". The corresponding results are presented in tables 5.12 and 5.13. The calculations of the statistical variables as a function of the substance class are shown in tables 9 to 14 in the appendix. The corresponding plots are largely identical to the previous ones and are therefore only shown in figures 3 and 4 in the appendix.

As expected, the results for the reoptimised PM6-PCM geometries in table 5.12 show a significant improvement over the models parameterised on the full Klicić data set. The previous best model, i.e. the individual PMV and pK_a correction using the NDDO potential (/B|P|N|I‖I|nt), also delivers good results here with an RMSE of 0.92. The corresponding reference model shows a comparable RMSE of 1.04, albeit on the full data set. It is noteworthy that the previously second best model with the same potential, a single PMV correction but an individual pK_a correction (/B|P|N|I‖I|nt), has an RMSE of 0.86, slightly better than the corresponding model with two individual corrections.

Overall, all models with the NDDO potential show similar results with the exception of the model parameterised according to the classical correction scheme from the original SAMPL6 publication, i.e. with a single parameter set in both correction models (/B|P|N|A‖A|nt). For this model, the exclusion of thiols only leads to an improvement of the RMSE from 2.31 to 2.24. Observation of the statistical values for the individual substance classes in table 5.10 and comparison with table 5.12 shows that even with the exclusion of the thiols, high deviations occur for the other substance classes that undergo a transition between the charge states 0 to -1, such as "acids" and "phenols". In addition, the substance classes "amines" and "heterocycles" also show very high deviations. It can therefore be assumed that this model is not capable of adequately describing other classes of substances in addition to thiols.

The remaining ESP variants give very similar results in combination with the classical parameterisation strategy, which uses only one set of parameters for each correction

Table 5.12.: Reparametrisation of the models from figure 5.8 and table 5.10 without the substance class "thiols" as defined in the Klicić data set. The models were obtained from ONIOM-EC-RISM/B@PFL single point calculations on PM6-PCM reoptimised structures. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 3 for the corresponding plots. Refer to table 5.4 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	1.22	1.00	0.00	0.90	0.53	0.90	0.64	-128.69
		/B|P|M|A‖A|nt	acids	1.17	0.96	-0.44	1.03	-0.55	0.85		
			bases	1.26	1.02	0.38	0.83	1.43	0.95		
	All	Individual	All	1.03	0.81	-0.00	0.93	0.37	0.93		
		/B|P|M|A‖I|nt	acids	0.98	0.81	-0.00	0.85	0.64	0.85	0.53	-106.10
			bases	1.07	0.81	-0.00	0.95	0.33	0.95	0.73	-148.02
	Individual	All	All	1.32	1.07	0.00	0.89	0.61	0.89	0.67	-133.99
		/B|P|M|I‖A|nt	acids	1.41	1.14	0.02	0.96	0.19	0.75		
			bases	1.23	1.00	-0.02	0.86	0.86	0.93		
	Individual	Individual	All	1.22	0.97	0.00	0.90	0.53	0.90		
		/B|P|M|I‖I|nt	acids	1.27	1.03	0.00	0.75	1.09	0.75	0.52	-104.13
			bases	1.18	0.91	0.00	0.93	0.40	0.93	0.73	-146.45
Exact (NDDO)	All	All	All	2.24	2.02	0.00	0.67	1.78	0.67	0.56	-110.12
		/B|P|N|A‖A|nt	acids	2.20	2.06	2.06	0.89	2.56	0.91		
			bases	2.28	1.99	-1.79	0.73	-0.15	0.96		
	All	Individual	All	0.86	0.68	-0.00	0.95	0.26	0.95		
		/B|P|N|A‖I|nt	acids	0.76	0.62	-0.00	0.91	0.39	0.91	0.57	-115.27
			bases	0.93	0.73	-0.00	0.96	0.25	0.96	0.73	-143.40
	Individual	All	All	0.99	0.82	0.00	0.93	0.35	0.93	0.67	-135.09
		/B|P|N|I‖A|nt	acids	0.94	0.77	-0.24	1.02	-0.35	0.89		
			bases	1.04	0.86	0.21	0.89	0.90	0.96		
	Individual	Individual	All	0.91	0.72	0.00	0.95	0.29	0.95		
		/B|P|N|I‖I|nt	acids	0.83	0.68	0.00	0.89	0.47	0.89	0.59	-117.11
			bases	0.97	0.75	0.00	0.96	0.27	0.96	0.73	-146.03
Point charge	All	All	All	1.23	1.02	0.00	0.90	0.54	0.90	0.57	-115.18
		/B|P|P|A‖A|nt	acids	1.37	1.18	-0.47	0.75	0.65	0.75		
			bases	1.09	0.87	0.41	0.90	1.03	0.95		
	All	Individual	All	1.13	0.90	-0.00	0.92	0.45	0.92		
		/B|P|P|A‖I|nt	acids	1.29	1.05	0.00	0.75	1.12	0.75	0.58	-114.92
			bases	0.98	0.76	-0.00	0.95	0.28	0.95	0.61	-122.92
	Individual	All	All	1.24	0.96	-0.00	0.90	0.54	0.90	0.59	-118.41
		/B|P|P|I‖A|nt	acids	1.46	1.19	-0.31	0.70	0.99	0.69		
			bases	1.01	0.76	0.27	0.92	0.75	0.96		
	Individual	Individual	All	1.20	0.88	0.00	0.91	0.51	0.91		
		/B|P|P|I‖I|nt	acids	1.42	1.12	-0.00	0.69	1.36	0.69	0.58	-115.84
			bases	0.96	0.67	0.00	0.96	0.27	0.96	0.61	-123.35
Exact (hr-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A‖A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (hr-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A‖A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

Table 5.13.: Reparametrisation of the models from figure 5.9 and table 5.11 without the substance class "thiols" as defined in the Klicić data set. The models were obtained from ONIOM-EC-RISM/B@PFL single point calculations on B3LYP/6-311+G**-PCM-optimised structures. See tables 5.6 and 5.7 for an explanation of the pK_a model IDs and figure 4 for the corresponding plots. Refer to table 5.5 for the PMV correction parameters used in the pK_a correction models presented in this table.

Potential	Charge fit	pK_a fit & ID	pK_a type	RMSE	MAE	MSE	m'	b'	R^2	m	b
Exact (Mult.)	All	All	All	1.20	0.95	-0.00	0.91	0.51	0.91	0.69	-140.17
		/B|B|M|A‖A|nt	acids	1.20	0.97	-0.56	0.82	0.23	0.83		
			bases	1.20	0.94	0.49	0.88	1.22	0.95		
	All	Individual	All	1.05	0.81	-0.00	0.93	0.39	0.93		
		/B|B|M|A‖I|nt	acids	1.06	0.84	-0.00	0.83	0.76	0.83	0.70	-140.80
			bases	1.05	0.78	-0.00	0.95	0.32	0.95	0.74	-151.84
	Individual	All	All	1.20	0.95	0.00	0.91	0.51	0.91	0.73	-147.80
		/B|B|M|I‖A|nt	acids	1.28	1.05	-0.15	0.77	0.85	0.76		
			bases	1.12	0.87	0.13	0.92	0.60	0.94		
	Individual	Individual	All	1.19	0.94	0.00	0.91	0.50	0.91		
		/B|B|M|I‖I|nt	acids	1.27	1.05	0.00	0.76	1.08	0.76	0.71	-144.31
			bases	1.11	0.85	-0.00	0.94	0.36	0.94	0.74	-150.91
Exact (NDDO)	All	All	All	2.34	2.09	0.00	0.64	1.93	0.64	0.57	-113.76
		/B|B|N|A‖A|nt	acids	2.36	2.12	2.11	0.73	3.30	0.84		
			bases	2.31	2.06	-1.84	0.74	-0.25	0.95		
	All	Individual	All	1.00	0.80	-0.00	0.93	0.35	0.93		
		/B|B|N|A‖I|nt	acids	1.01	0.85	0.00	0.84	0.69	0.84	0.66	-134.60
			bases	0.98	0.76	-0.00	0.95	0.28	0.95	0.73	-145.62
	Individual	All	All	1.13	0.91	-0.00	0.91	0.46	0.91	0.69	-138.81
		/B|B|N|I‖A|nt	acids	1.16	0.95	-0.50	0.84	0.22	0.83		
			bases	1.11	0.87	0.44	0.89	1.09	0.95		
	Individual	Individual	All	1.01	0.82	0.00	0.93	0.36	0.93		
		/B|B|N|I‖I|nt	acids	1.04	0.86	0.00	0.83	0.73	0.83	0.69	-138.00
			bases	0.99	0.78	-0.00	0.95	0.28	0.95	0.73	-148.93
Point charge	All	All	All	1.34	1.00	0.00	0.88	0.63	0.88	0.60	-121.85
		/B|B|P|A‖A|nt	acids	1.57	1.20	-0.48	0.54	1.55	0.69		
			bases	1.09	0.84	0.42	0.93	0.84	0.95		
	All	Individual	All	1.21	0.94	-0.00	0.90	0.52	0.90		
		/B|B|P|A‖I|nt	acids	1.42	1.15	-0.00	0.69	1.35	0.69	0.78	-157.58
			bases	1.00	0.76	-0.00	0.95	0.29	0.95	0.62	-125.16
	Individual	All	All	1.35	0.99	0.00	0.88	0.64	0.88	0.61	-124.19
		/B|B|P|I‖A|nt	acids	1.64	1.22	-0.40	0.50	1.81	0.65		
			bases	1.02	0.78	0.35	0.95	0.68	0.96		
	Individual	Individual	All	1.25	0.94	-0.00	0.90	0.55	0.90		
		/B|B|P|I‖I|nt	acids	1.52	1.19	-0.00	0.65	1.55	0.65	0.80	-162.22
			bases	0.96	0.71	-0.00	0.96	0.27	0.96	0.62	-125.94
Exact (*hr*-ref.)[24]	All	All	All	1.04	0.87	-	-	-	0.98	0.74	-150.72
		hr|B|E|A‖A	acids	0.93	0.77	-	-	-	0.93		
			bases	1.14	0.97	-	-	-	0.95		
Point charge (*hr*-ref.)[24]	All	All	All	1.88	1.34	-	-	-	0.92	0.62	-126.03
		hr|B|P|A‖A	acids	2.04	1.33	-	-	-	0.69		
			bases	1.70	1.35	-	-	-	0.92		

(/B|P|M|A‖A|nt and /B|P|P|A‖A|nt). At the same time, the same picture emerges as for the NDDO based models: The models with a single PMV correction and an individual pK_a correction show slightly better results than the corresponding model with two individual corrections.

Due to the exclusion of thiols from the training data set, the pK_a correction parameters change significantly and approach the respective *hr* reference models. In the case of separate parameterisation for acids and bases, the parameters of the acids approach the second parameter set. In all cases this leads to an increase in m.

This effect is much more pronounced for the models parameterised on the original B3LYP-PCM geometries. The corresponding results are shown in table 5.13, figure 4 and in the appendix in tables 12 to 14. Especially for the multipole ESP models the differences for the parameter m are small. However, larger deviations are observed for the second parameter b. It also deviates more from the *hr* reference parameters for acids, so that it can be assumed that the ONIOM model for acids does not extrapolate to the corresponding ΔG_i of the reference method.

As before for the complete Klicić data set, the deviations from the experimental reference for the NDDO and point-charged models are also smaller for the PM6-PCM structures than for the B3LYP PCM structures on the data set without thiols. However, good results with similar statistical parameters to the respective *hr* reference models can also be obtained for the models with individual correction for acids and bases.

In summary, it was demonstrated that the quality of predictions using the classical correction approach, which employs a single set of parameters for all charge states in the PMV correction and a single set of parameters for the pK_a correction, could be enhanced by applying individual corrections. It was also argued that predictions for the "acid" substance class, especially thiols, are responsible for the lower overall prediction quality. Parameterisations omitting this substance class yield improved results. However, as the predictive quality of the models has only been assessed using the same data set on which they were parameterised, these findings should not be taken as a final statement on their accuracy, especially as the individual corrections may be overfitted. As mentioned above, the actual predictive power of the models will be validated on an independent test data set in the following sections.

5.3. Acidity constant prediction for the SAMPL6 data set

The models parameterised in the previous sections are tested using the pK_a data set from the SAMPL6 challenge. Originally constructed as part of a blind prediction challenge, this dataset consists of 24 small molecules that resemble kinase inhibitor fragments and has been used to assess the performance of current prediction methods.[25] As described in the previous sections, submissions to the SAMPL6 challenge were made using an *hr*-EC-RISM model,[24,25] which can be used in the context of this work as a reference

model in addition to direct comparison with later published experimental pK_a values.[25]

5.3.1. Graph-based transfer of ONIOM partitions

Until now, the size of the data sets used to parameterise the PMV and pK_a corrections has prevented the partitioning of their molecules. To circumvent this problem, the partition-free limit (PFL) has been applied to the ONIOM-EC-RISM schemes, which also ensures that the resulting models are free from any ONIOM partitioning error. Nevertheless, it should be emphasised that in most cases the application of the PFL increases the computational cost, and therefore the aim of the following sections is to evaluate whether the application of the PFL parameterisations can be applied to ONIOM-EC-RISM calculations on partitioned molecules and whether they are a suitable approximation to the unavailable parameters that would in principle be obtained from calculations on partitioned molecules. In principle, the SAMPL6 dataset is an ideal test case in this respect, as it contains only a small number of molecules and therefore allows manual partitioning.

However, some of the SAMPL6 molecules are multiprotic species and can occur in a variety of tautomeric forms, which leads to additional considerations for the ONIOM calculations: In a two-layer ONIOM-EC-RISM calculation, partitions are defined by specifying the subset of atoms that are treated at the *high* level of theory and the corresponding link atoms by specifying the link atom host and link atom connection. For practical reasons this is done by specifying the corresponding atoms by their indices as defined by the atom order in the input coordinate file. As a consequence, the indices of two protonation states of the same molecule may differ in terms of the indices that need to be specified to obtain an identical partition, thus increasing the number of structures that need to be manually partitioned.

This is illustrated by an example: Consider two protonation states of a hypothetical molecule AH-B-R, where A and B are protonation centres actively involved in the protonation processes. These processes can be tautomerisation (AH-B-R $\rightleftharpoons$ A-BH-R) or (de)protonation reactions (AH-B-R $\rightleftharpoons$ [A-B-R]$^-$ + H$^+$). It is assumed that for this molecule a partitioning is given where both protonation centres A and B and the proton are included in the *model* system and that the residue R is only described at a *low* level of theory. If this partition is specified for AH-B-R by indices and is to be transferred to its tautomer A-BH-R, great care must be taken to ensure that the respective indices remain unchanged.

The same applies to protonation and deprotonation reactions. If the set of indices is to be specified once and reused for other protonation steps, the input coordinates must be ordered in such a way that the specified indices continue to point to the protonation centres A and B and the protons involved. This usually involves a great deal of manual work and is prone to error.

To overcome this problem, a method has been developed to automatically transfer

partitions from one protonation pattern of a molecule to another. As a result, for the SAMPL6 data set only 24 partitions need to be manually specified, one for each molecule, instead of one for each tautomer in each protonation state for each molecule, which quickly leads to a combinatorial explosion.

The process is based on a few simple ideas. First, the transfer of a partition involves matching the indices of one structure to the indices of the second structure. Second, given two structures that differ only in their protonation pattern, their maximum common substructure (MCS) excludes all the protons involved in the protomeric reaction that transforms one structure into the other. Third, the MCS graphs of the two structures can be considered isomorphic, i.e. they differ only in their labelling, while their connectivity is identical. Therefore, one can create a mapping from one set of indices to the other structure's set of indices by matching the indices of their MCS graphs.

The graph-based transfer of the ONIOM partitions presented here can therefore be divided into three steps. First, the set of protons involved in the reaction is identified by calculating the MCS of both structures. Next, the isomorphism of the two MCS graphs is computed, which is then used in a final step to map the indices specifying a partition for one structure to the corresponding set of indices encoding the same partition for the second structure.

The first step is initiated by reading both structures from either Protein Data Bank (PDB) files or AMBER format files. In addition, the indices specifying the partition for one structure are read from simple text files. The associated structure is referred to as the reference structure, while the second structure for which a corresponding set of indices must be found is referred to as the transfer structure.

The MCS of both structures is then calculated using the fMCS algorithm as implemented in RDKit. The result is a canonical SMILES string that represents the MCS, but no longer carries information about the associated indices and therefore the indices of the excluded protons. This information must be recovered in the following steps. This can be done by iteratively excluding protons from both the reference and transfer structure and comparing the resulting canonical SMILES strings with the MCS SMILES until both matching strings are found. The number of protons to be excluded can be easily obtained as the difference between the respective sizes of the input structures and the MCS.

Since the indices of the protons that differ between the reference and the transfer structure are now known, all other indices are part of the MCS and can be used in the next step to generate the isomorphism between the two MCS graphs. Note that the iterative SMILES comparison is only used because the fMCS algorithm returns a SMILES string. In future revisions of this partition transfer algorithm, the iterative comparison can be avoided by replacing the fMCS algorithm with another implementation that gives a more suitable output. However, for the purpose of this work and the sizes of the SAMPL6 molecules, the additional computational effort caused by the SMILES comparison is acceptable.

The next step is to compute the isomorphism of the two structures. This is done by constructing their graphs from the connectivity given in the input files. In order to exclude the protons that were identified in the first step, they are disconnected from their respective graphs. If the total number of protons in the two structures is different, disconnected vertices are added to the smaller structure until the number of vertices is equal to that of the other structure. Disconnecting the vertices representing the protons that differ between the structures ensures that the remaining connected vertices represent the MCS. In addition, the inclusion of these protons as disconnected vertices ensures that the indices of all vertices match those of their respective input structures.

The graphs constructed in this way have the same number of vertices, edges and connectivity. Thus, they differ only in the labels assigned to the vertices, i.e. the indices of the respective input structure, and their isomorphism can be used to transfer the partitioning from the reference to the transfer structure. This isomorphism is computed using the "nauty" algorithm with the "dreadnaut" interface,[123] by computing the canonical labelling of both graphs. This labelling is identical for all isomorphic graphs and is similar in concept to the canonical labelling used in the context of SMILES strings. Since there is a mapping from any graph labelling to the canonical labelling and vice versa, the isomorphism mapping from the reference to the transfer graph can be constructed by composition. This isomorphism is computed directly by the nauty algorithm and can then be used in a subsequent step to transfer the partition to the transfer structure. Protons that differ between structures are automatically added to the *model* system.

See the electronic supplementary material for the code of the partition transfer algorithm.[116]

5.3.2. Partitioning of the SAMPL6 structures

The molecules in the SAMPL6 dataset were partitioned to include all protonation centres in the *model* system. Protonation centres were identified by comparing all macro and micro states provided by the challenge organisers.[25] All atoms that showed a change in protonation from one state to another were marked as protonation centres. Whenever possible, molecules were split by cutting single bonds, as this is considered the safest way to avoid large partitioning errors.[5] Cuts were chosen so that all protonation centres are contained in a single, connected fragment. As a general rule, partitions were preferred that resulted in chemically reasonable *model* systems. This may of course be subject to bias.

All atoms involved in the protonation process were added to the *model* system to ensure that the zero point of the ONIOM energy scale remains identical for all microstates. Two cases have to be considered: First, relative energies where the composition of both states is identical. In this case the composition of the *model* system must be identical for both states, i.e. all Hamiltonians remain unchanged. For the protonation problems we consider later, this means that the *model* system must be chosen in such a way that

the tautomeric proton cannot cross the ONIOM boundary. Secondly, when calculating relative energies between states where the composition of the system is changed, all changes must be modelled within the *model* system, i.e. all changes are modelled in all Hamiltonians. For example, if the system undergoes protonation or deprotonation, this condition is met if the protonation site is within the *model* system. If the deprotonation or protonation process takes place outside the *model* system and its composition remains unchanged, the zero points of the energy scales for both states are not identical and therefore do not cancel in their relative energies.

All partitions constructed in this way are shown in figure 5.10. A special subset of molecules is given by SM02, SM04, SM07, SM09, SM12 and SM13, all of which share a common scaffold shown in figure 5.11. This scaffold consists of a heterocyclic aromatic ring system containing two endocyclic nitrogen atoms and an exocyclic amine, which have been identified as protonation centres. All these molecules differ in the substituent position R_1 of the exocyclic amine as well as in two other exocyclic positions R_2 and R_3. Since these substituents do not contain protonation centres, a suitable *model* system can be constructed from their common scaffold by cutting the bonds with residues R_1 to R_3 and saturating the open valencies with hydrogen link atoms. With the exception of SM13, for all these molecules R_2 and R_3 are hydrogen atoms, so only the bond to R_1 needs to be cut. In contrast, SM13 shows two methoxy groups at positions R_2 and R_3. The exclusion of these groups from the *model* system means that two polarised carbon-oxygen bonds must be cleaved. This is accepted as it provides an opportunity to test the robustness of the ONIOM-EC-RISM schemes with respect to more drastic partition schemes.

The common scaffold of SM06 and SM22 shown in figure 5.11 also shows similarities to the previously discussed structures. However, the arrangement of the protonation centres in SM06 and SM22 allows only for a limited number of partitions that do not result in cuts through conjugated ring systems or that would separate the protonation centres into unconnected fragments. Therefore, for these molecules, a *model* system was chosen that cuts all halogen bonds, as these were identified as the only reasonable cuts. The remaining structures, with the exception of SM14 and SM15, do not show pronounced similarities, but their partitioning problems are drastically different and therefore their and all the remaining partitionings will be discussed individually.

SM01 shows three protonation centres connected by an aromatic ring system that should not be cut in order to keep the centres in one fragment and to avoid cutting conjugated bonds. Therefore, the alkyl chain of the lactam ring system was excluded from the *model* system.

SM03 contains a carbon protonation centre which reduces the number of single bonds that can be cleaved. In fact, the distribution of all protonation centres allows only two single bonds to be cleaved, i.e. the cuts that separate the phenyl and thiophene moieties from the core fragment.

The allowed cuts for SM05, which keep the protonation centres in one connected

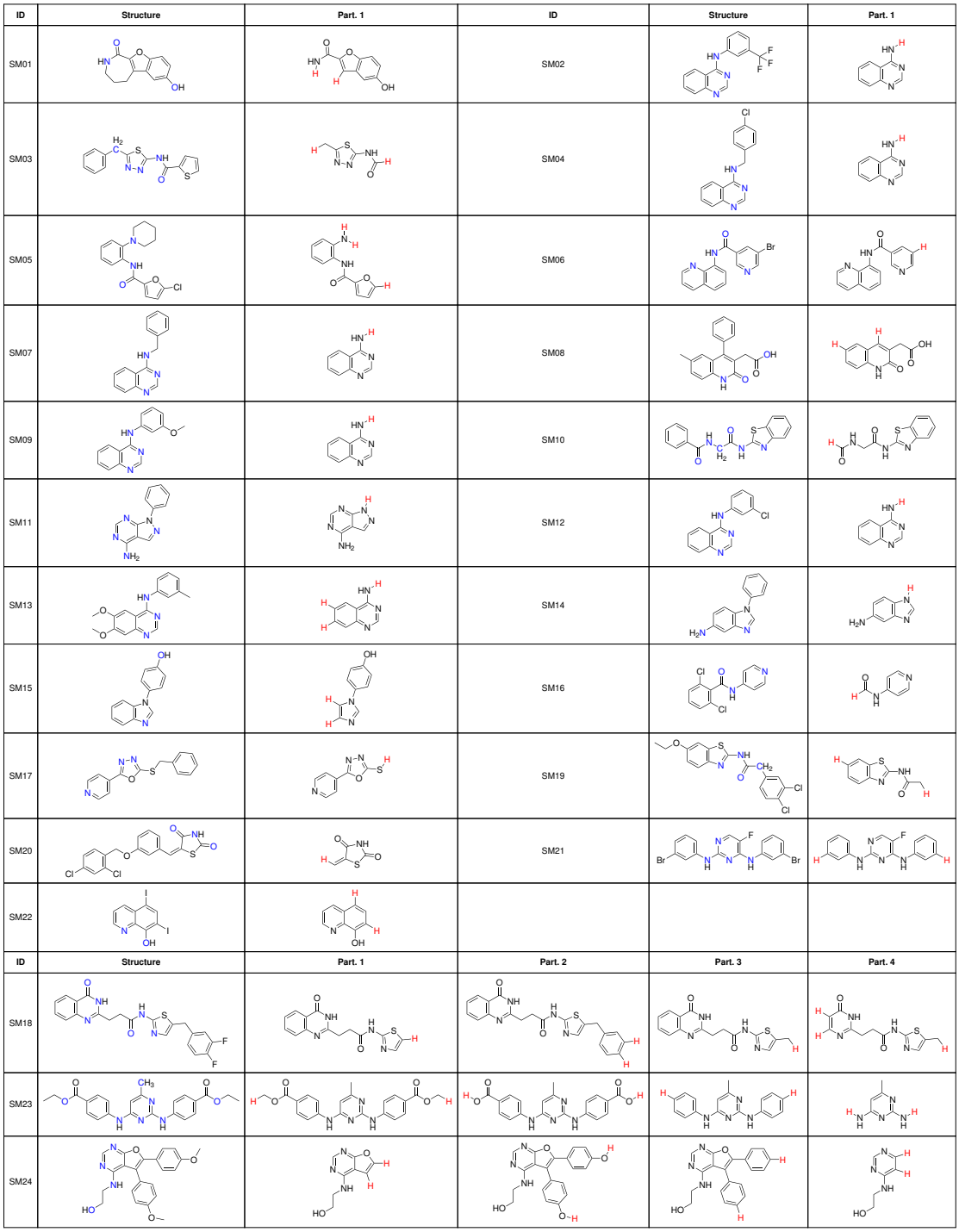

Figure 5.10.: Partitions of the SAMPL6 data set. Protonation centres as defined by the challenge organisers are highlighted in blue, while link atoms used to saturate open valences of the *model* systems are marked in red. For SM18, SM23 and SM24 multiple partitions were created. pK_a values and statistical parameters reported were calculated based on the smallest available *model* systems.

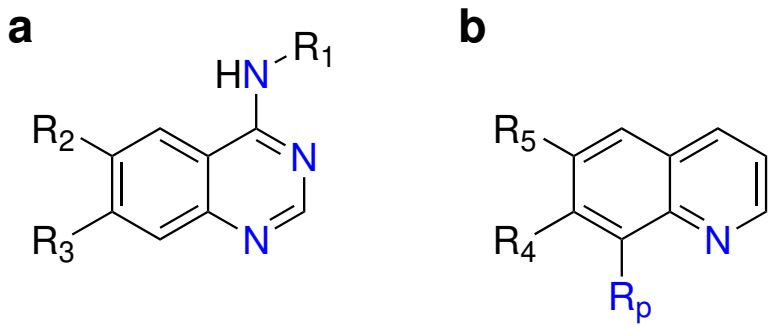

Figure 5.11.: Common scaffolds for molecules SM02, SM04, SM07, SM09, SM12, SM13 (a) or SM06 and SM22 (b). Atoms that are deemed to be involved in the protonation processes by the SAMPL6 challenge organisers are highlighted in blue. Note that R_p also includes protonation centres for SM06. The respective ONIOM *model* systems are constructed by replacing the substituent positions marked by R_1 to R_5 with hydrogen link atoms.

fragment, would require a cut within the piperidine moiety. The second set of single bonds that can be cut connects the furan moiety to the rest of the molecule and to the halogen bond of the furan group. The latter was chosen because the former would mean cutting through a conjugated system. A more conservative partitioning scheme was therefore chosen in this case.

The partitioning of SM08 is more trivial as there are only two single bonds which do not split the protonation centres into multiple fragments. Therefore, the bonds to the methyl substituent of the aromatic ring system and to the phenyl group are cleaved.

SM10, SM11, SM14 and SM16 show a similar partitioning problem. The only suitable bonds that can be cut are the single bonds to the phenyl groups.

Although SM15 is structurally similar to SM14, the former has an additional hydroxyl substituent on the phenyl group which has been identified as a protonation centre. This prohibits the exclusion of this group from the *model* system by the partitioning rules defined above. This also means that the only cuts that are remotely suitable for these systems are those that join the two six- and five-membered rings. Although it is generally advised not to cut aromatic systems, this partition is therefore used as a test case for more drastic partitionings, similar to SM13.

In the case of SM17, there are several suitable *model* systems that can be constructed. Here it was decided to exclude the benzyl group and include the sulphur atom in the *model* system as it was expected to affect the nitrogen protonation centres in the vicinity.

Similar to SM03, SM19 has a carbon protonation centre. Therefore, the benzyl group cannot be excluded, but the cut can be moved to the nearest single bond, thus excluding the chlorosubstituted phenyl group. At the other end of the molecule, the ethoxy group

was cleaved from the core fragment.

In SM20, all the protonation centres are in close proximity to each other and can be isolated as a single contiguous fragment by a variety of cuts. Here a large part of the molecule can be excluded from the expensive *high* level calculation by cutting the carbon-carbon bond connecting the substituted phenyl group to the fragment containing all the protonation centres. This was considered an appropriate cut as this bond is expected to be unpolarised, and it moves the ONIOM boundary far enough away from the protonation centres while reducing the size of the *model* system to a minimum that is still chemically reasonable.

To reduce the computational cost of SM21, it was decided to remove the electron-rich bromide atoms, while including the fluoride atom close to the four nitrogen protonation centres, in order to model the electron-withdrawing effect on the ring system at the *high* level of theory.

The sizes of SM18, SM23, SM24 allow the definition of multiple partitions and are therefore used as a test case to investigate the influence of different *model* system sizes. For example, the symmetry of the substituent groups of the core heterocyclic system of SM23 allows a symmetric reduction of the *model* system size. Starting with the first partition P1, only the terminal methyl group is removed from the rest of the molecule. In P2 the adjacent methylene group is also removed. The removed ethoxy group is part of an ether functional group where the oxygen atom has been identified as the protonation centre, thus limiting the number of possible *model* systems. However, it is expected that the protonated states of the ethyl ester group are not highly populated and can be excluded from the pK_a calculation without significantly affecting the prediction quality. For this reason it is possible to exclude the ester functional group from the *model* system and create the additional partitions 3 and 4, in which the *model* system has been drastically reduced to the core fragment.

In contrast, for SM24 no additional assumption need to be made as no protonation centres need to be excluded from the *model* system and a similar stepwise reduction of the *model* system can be achieved by excluding the phenolic group or parts of it. To obtain the smallest *model* system, the bonds between the six-membered heterocycle and the five-membered heterocycle are cut.

A similar partitioning scheme was used for the smallest *model* system, SM18. This molecule has two fused rings, only one of which contains a protonation centre, so the second ring can be excluded by cutting the connecting bonds. As before, this can be used to evaluate the robustness of the ONIOM method with respect to more drastic partitioning schemes.

Note that the numbering of the partitions does not necessarily reflect an increase or decrease in the size of the *model* system, but rather the order in which the partitions were created. For all three molecules, the fourth partition represents the smallest *model* system.

For molecules where multiple partitions were generated, the overall model quality for

the SAMPL6 dataset was evaluated using the pK_a prediction obtained from the smallest available *model* system, unless otherwise specified. More specifically, this is P4 for the set of molecules, where multiple partitions were created. Given the expectation that inclusion of larger parts of the molecule in the *model* system will lead to improved results, the results given can thus be considered as a lower bound for the prediction quality.

5.3.3. Computational details

As the partition-free limit was not used for the pK_a predictions on the SAMPL6 data set, all relevant ONIOM-EC-RISM schemes /A, /B and /X were applied to obtain single point energies. This also means that in addition to the conformer optimisation strategy used for the previous two data sets, i.e. the use of the original B3LYP-PCM and the reoptimised PM6-PCM geometries, a set of reoptimised ONIOM-PCM/X geometries could be obtained. These conformers were generated using the ONIOM-PCM implementation of Gaussian16 and the partitions specified and transferred above, starting from the original B3LYP-PCM structures. Only the /C and /X implementations in Gaussian16 allow geometry optimisations. The latter was chosen here because it was considered less approximate.

All calculations were performed using MP2/6-311+G** as the *high* level of theory and PM6 as the *low* level of theory. Three sets of energies were calculated by applying either the multipole approximation of the ESP, the NDDO-ESP or the point-charge based ESP to the 3D-RISM solver. As described in section 4.6, link atoms were placed using the link bond scaling approach (equation 2.8).

As for the calculations on the MNSOL and Klicić datasets, the settings for the EC-RISM calculations were identical to those described in the SAMPL6 publication.[24] In all cases, the PSE-2 closure and a 128^3 grid with a 0.3 Å grid spacing were used for the single point calculations. The ORCA calculations were performed using the "TightSCF" convergence criteria. The sub-calculations performed with EMPIRE were started with an energy convergence threshold of 10^{-6} kcal mol^{-1} and 10^{-6} for the convergence of the maximum off-diagonal CFC element.

To ensure an identical evaluation of the models compared to the *hr* reference models, the Mathematica notebooks used in the SAMPL6 challenge were provided by Nicolas Tielker. These notebooks were then modified to allow the use of the ONIOM-EC-RISM output and the new ONIOM-based PMV and pK_a corrections. The associated raw data, such as atomic coordinate files, can be found in the electronic supplementary material.[116]

5.3.4. Initial model selection with ONIOM-EC-RISM/B

In the last few sections a large number of models have been fitted to circumvent some of the difficulties observed in predicting solvation free energies and pK_a values for the MNSOL and Klicić data sets. However, the combination of all PMV and pK_a correction parameters with all ESP and ONIOM approximations leads to a large number of final pK_a models, the manual evaluation of which would require a considerable effort. However, based on the results of the previous chapters, it can also be assumed that not all of these models will lead to accurate predictions, so a selection of promising models has to be made as a first step.

Sections 5.3.8 and 5.3.9 provide a summary of the best performing pK_a prediction models and an assessment of the speed-up gained through ONIOM-EC-RISM.

The preselection of pK_a predicion models will first be done using the ONIOM-EC-RISM/B model, since the previous parameterisations were also carried out at the /B@-PFL level of theory. Subsequently, these pre-selected promising models can be evaluated at the level of theory of the remaining ONIOM-EC-RISM approximations, drastically reducing the overall required effort. It is expected that models that perform well with the /B scheme will also perform well with the other methodologically similar ONIOM-EC-RISM schemes.

A first pre-selection of models can be made based on the structure of the SAMPL6 data set, which includes macrostates with total charges between -3 and +4, whereas the Klicić data set used to parameterise the pK_a correction contains only single positively and negatively charged molecules and neutral molecules, but no multiple charged species. The individual pK_a correction, where acids, defined as the transition between charge states 0 and -1, and bases, defined as the transition from +1 to 0, are corrected with a separate set of parameters, cannot be trivially applied to the SAMPL6 dataset. This separate correction was originally used to avoid the poor prediction quality of anions, which in turn affects the prediction quality of the acids. If this model is to be applied to data sets containing multiply charged species, it is necessary to decide which set of parameters should be applied to which transitions. For example, if one wants to apply this model to the SAMPL6 dataset, it needs to be decided whether to apply the base set of parameters to transitions between positive charge states, and whether to apply the acid set of parameters to the problematic 0 to -1 transitions only, or to all transitions between negative charge states.

In the previous chapter, however, it could be seen that the application of the individual PMV correction, in which the individual charge levels were parameterised separately, leads to very similar results to the individual pK_a correction discussed here. Therefore, the additional effort and ambiguity resulting from the application of the individual pK_a correction will be avoided and only the classical correction approach with a single set of parameters m and b for all transitions will be used.

Therefore, combinations of the following parameters are used: Firstly, for the PMV

Table 5.14.: Overview of all mixed models resulting from the application of PMV and pK_a corrections that were not parameterised at the same level of theory. Here, PMV correction models parameterised in section 5.1 are combined with the best performing pK_a correction models from the original SAMPL6 publication[24], for the respective ESP approximation, i.e. "exact" or "point charge". Note that all *hr*-MP2 pK_a parameters taken from the original SAMPL6 publication were obtained from B3LYP-PCM geometries. The last rows define the model IDs for these original SAMPL6 models. See tables 5.1 and 5.6 for a definition of the model IDs for the PMV correction, as well as other pK_a corrections tested in this chapter. See the main text for an explanation of the "Indiv,+" PMV correction approach. Tables showing the associated PMV correction parameters are reported next to the PMV model ID. The pK_a correction parameters of the *hr*-reference model can be found alongside the newly parameterised models, e.g. in table 5.11.

PMV correction model						pK_a correction model		Mixed model ID
Optimisation	EC-RISM	Potential	Charge fit	Model ID	Tab.	EC-RISM	Model ID	
B3LYP-PCM	/B@PFL	Exact (Mult.)	All	/B|B|M|A	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|M|A‖*hr*|A
			Indiv.	/B|B|M|I	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|M|I‖*hr*|A
			Indiv.,+	/B|B|M|I+	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|M|I+‖*hr*|A
	/B@PFL	Exact (NDDO)	All	/B|B|N|A	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|N|A‖*hr*|A
			Indiv.	/B|B|N|I	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|N|I‖*hr*|A
			Indiv.,+	/B|B|N|I+	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|N|I+‖*hr*|A
	/B@PFL	Point charge	All	/B|B|P|A	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|P|A‖*hr*|A
			Indiv.	/B|B|P|I	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|P|I‖*hr*|A
			Indiv.,+	/B|B|P|I+	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	/B|B|P|I+‖*hr*|A
PM6-PCM	/B@PFL	Exact (Mult.)	All	/B|P|M|A	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|M|A‖*hr*|A
			Indiv.	/B|P|M|I	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|M|I‖*hr*|A
			Indiv.,+	/B|P|M|I+	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|M|I+‖*hr*|A
	/B@PFL	Exact (NDDO)	All	/B|P|N|A	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|N|A‖*hr*|A
			Indiv.	/B|P|N|I	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|N|I‖*hr*|A
			Indiv.,+	/B|P|N|I+	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|N|I+‖*hr*|A
	/B@PFL	Point charge	All	/B|P|P|A	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|P|A‖*hr*|A
			Indiv.	/B|P|P|I	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|P|I‖*hr*|A
			Indiv.,+	/B|P|P|I+	(5.4)	*hr*-MP2	*hr*|B|E|A‖A	/B|P|P|I+‖*hr*|A
B3LYP-PCM	*hr*-MP2	Exact[24]	All[24]	*hr*|B|E|A	(5.5)	*hr*-MP2	*hr*|B|E|A‖A	-
		Point charge[24]	All[24]	*hr*|B|P|A	(5.5)	*hr*-MP2	*hr*|B|P|A‖A	-

correction, the *hr*-MP2 correction parameters from the original SAMPL6 paper[24] can be used. Secondly, the newly parameterised corrections from the PFL application can be used, namely the correction with a single set of parameters for all charge states ("/B@PFL (All)") and the corresponding correction with an individual correction ("/B@-PFL (indiv.)"). In addition to these models, an individual charge correction scheme is tested where the parameters obtained for the cationic subset of the MNSOL are also applied to all negatively charged species. This was done to test if the application of a set of parameters, unaffected by the problematic prediction quality for the anionic

species, to all charged species can improve the overall results. This correction scheme is abbreviated as "/B@PFL (indiv.,+)" and marked in the model IDs as "I+".

These PMV correction schemes must then be combined with a pK_a correction scheme. Three options are available. The *hr*-MP2 parameters from the SAMPL6 paper and the newly parameterised corrections at the /B@PFL level. There are two options for the latter. One where the full Klicić dataset was used for the parameterisation and one where the model was fitted excluding thiols from the dataset as detailed previously.

The final pK_a correction model used in conjunction with the free energies predicted according to the /A, /B or /Xa scheme is defined by a set of four parameters, c_V, c_q, m and b. For /Xb, an additional set of PMV parameters is required to account for the reversed order of operation of the ONIOM extrapolation and PMV correction, as described in section 4.6.4. Thus, all models are fully defined by the specification of the correction models, e.g. by their model IDs as shown in table 5.6, the underlying ONIOM-EC-RISM scheme and the set of geometries used for the single point calculations.

In contrast to the pK_a correction models parameterised at the /B@PFL level of theory (table 5.6), the combination of a /B@PFL set of PMV parameters with the *hr*-MP2 pK_a correction parameters from the original SAMPL6 paper results in a mixed model where the pK_a correction is not parameterised at the level of the PMV correction. The model IDs of these mixed models are defined in table 5.14. /Xb models are defined later in tables 5.22 and 5.23.

For example, the model associated with ID /B|B|M|A‖*hr*|A is based on the PMV correction /B|B|M|A. As this PMV correction is obtained from /B@PFL calculations on B3LYP-PCM geometries, the associated parameters can be found in table 5.5, which contains all models obtained at this level and these geometries. The *hr*-MP2 parameters used for the pK_a correction can be found in each table in section 5.2. To stress that the pK_a correction was derived at a different level of theory than the PMV correction, the level of theory of the pK_a correction is added as a identifier to the model ID, e.g. /B|B|M|A‖*hr*|A shows that the *hr*-pKa correction was applied to free energies that were corrected at the /B@PFL level. In contrast, for /B|B|M|A|||A both corrections were derived using the same level of theory and the standard model IDs, as introduced during the last chapter are applied.

Examples on how to retrieve the associated model parameters are also provided in the previous sections dealing with the parameterisation of the PMV and pK_a correction.

B3LYP-PCM geometries

The analysis will begin by examining the results obtained from the B3LYP-PCM-optimised geometries and then move on to the more approximate levels of optimisation. The statistical analysis for the three pK_a correction schemes is shown in tables 5.15. The corresponding predicted pK_a values are depicted in figures 5.12, 5.13 and 5.14, while their numerical values are given in tables 16 to 24 in the appendix. Note that all these

quantities have been calculated for the respective smallest *model* system available.

Note that the parameters m' and b' were obtained using the equation $\mathrm{p}K_{\mathrm{a,corr}} = m'\mathrm{p}K_{\mathrm{a,exp}} + b'$, as this is the procedure employed in the reference publication by Tielker et al.[24] for the SAMPL6 dataset. However, it is more common to use the equation $\mathrm{p}K_{\mathrm{a,exp}} = m''\mathrm{p}K_{\mathrm{a,corr}} + b''$. Nevertheless, the definition used here can still be employed alongside the other prediction measures to assess the model's performance, since an ideal model would also yield $m' = 1$ and $b' = 0$ for the regression equation used here, and any deviations would indicate that the model is less than ideal.

In addition, the *hr*-MP2 model from the SAMPL6 paper, which gave the smallest RMSE of 1.13, is shown in the tables. A suitable ONIOM-EC-RISM model, capable of extrapolating to the *hr*-reference, must therefore provide similar statistical values to this model.

As before, the ideal model must result in an MSE of zero and minimise the RMSE and MAE, while the parameters of the descriptive regression must be $m' = 1$, $b' = 0$ and $R^2 = 1$. The results of the *hr*-MP2 correction from table 5.15 clearly show that most models based on the NDDO potential do not allow an adequate $\mathrm{p}K_{\mathrm{a}}$ prediction. The application of the *hr*-MP2 PMV correction (*hr*|B|E|A‖A) leads to a high RMSE of 2.96. The application of the re-parameterised /B@PFL-PMV correction with a single parameter set (/B|B|N|A‖*hr*|A) does not lead to an accurate model either. Here the RMSE deteriorates to a value of 3.69. It is noticeable that both the MSE and b' have significantly lower values compared to the previous model. From these values and figure 5.12 it can be assumed that the experimental $\mathrm{p}K_{\mathrm{a}}$ values are significantly underestimated by the prediction models. Based on the figure and the slope parameter m', this is mainly due to the underprediction of the experimental $\mathrm{p}K_{\mathrm{a}}$ values between 0 and 5.

A reduction in the RMSE to a value of 3.47 can be achieved by using the individual PMV correction (/B|B|N|I‖*hr*|A). However, the results in table 5.15 show that the predicted values of this model are highly scattered around the line of the descriptive regression, which is also illustrated by the low coefficient of determination of 0.23. It can be assumed that the slight improvement in the RMSE is due to a random effect, but that it cannot be used for accurate prediction due to the high scatter of the calculated $\mathrm{p}K_{\mathrm{a}}$ values.

This effect can be eliminated by applying the PMV parameter set for the subset of cationic molecules to all ionic macrostates of the SAMPL6 dataset (/B|B|N|I+‖*hr*|A). This "/B@PFL (indiv.,+)" PMV correction drastically reduces the RMSE to a value of 1.59. The previously observed underestimation of the experimental values can also be significantly improved, resulting in an improved MSE of -0.37. The resulting RMSE is therefore only slightly larger than the more expensive EC-RISM reference calculations at the *hr*-MP2 level of theory. Since this correction approach simply replaces the parameters of the anionic species with those of the cationic species, the previous erroneous predictions can now be attributed to these anionic parameters.

At the /B@PFL level of theory, using the NDDO-ESP approximation, it is therefore

Table 5.15.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B single point calculations on B3LYP-PCM-optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.5, and pK_a correction parameters can be found in table 5.11 or table 5.13 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.86	1.69	-1.68	0.97	-1.49	0.91
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	0.85	0.71	-0.16	0.93	0.29	0.90
	/B@PFL (Indiv.)	*hr*-MP2	/B|B|M|I‖*hr*|A	2.60	1.94	-1.58	0.75	-0.09	0.52
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|M|I+‖*hr*|A	2.83	2.25	-2.09	0.50	0.94	0.50
Exact (NDDO)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	2.96	2.60	-2.04	1.41	-4.53	0.81
	/B@PFL (All)	*hr*-MP2	/B|B|N|A‖*hr*|A	3.69	3.34	-3.06	1.38	-5.35	0.81
	/B@PFL (Indiv.)	*hr*-MP2	/B|B|N|I‖*hr*|A	3.47	2.59	-2.35	0.41	1.21	0.23
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.59	1.32	-0.37	1.10	-0.99	0.79
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	1.38	1.13	-0.48	0.89	0.19	0.78
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	1.81	1.55	-1.25	0.86	-0.39	0.77
	/B@PFL (Indiv.)	*hr*-MP2	/B|B|P|I‖*hr*|A	2.62	2.14	-1.88	0.71	-0.14	0.58
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|P|I+‖*hr*|A	2.95	2.40	-2.30	0.53	0.55	0.53
Exact (Mult.)	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.13	0.89	-0.61	0.73	0.99	0.91
	/B@PFL (Indiv.)	/B@PFL	/B|B|M|I‖A	2.58	1.97	-1.76	0.59	0.71	0.52
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|M|I+‖A	2.91	2.32	-2.17	0.39	1.51	0.50
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|B|N|A‖A	2.28	2.08	-1.83	1.05	-2.15	0.81
	/B@PFL (Indiv.)	/B@PFL	/B|B|N|I‖A	2.72	1.85	-1.28	0.31	2.86	0.23
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.26	0.93	0.23	0.84	1.18	0.79
Point charge	/B@PFL (All)	/B@PFL	/B|B|P|A‖A	1.62	1.39	-0.99	0.76	0.49	0.77
	/B@PFL (Indiv.)	/B@PFL	/B|B|P|I‖A	2.35	1.83	-1.55	0.63	0.70	0.58
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|P|I+‖A	2.68	2.10	-1.92	0.46	1.31	0.53
Exact (Mult.)	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	1.39	1.20	-1.11	0.86	-0.25	0.91
	/B@PFL (Indiv.)	/B@PFL, n.t.	/B|B|M|I‖A|nt	3.13	2.54	-2.42	0.70	-0.60	0.52
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|M|I+‖A|nt	3.47	2.92	-2.90	0.46	0.35	0.50
Exact (NDDO)	/B@PFL (All)	/B@PFL, n.t.	/B|B|N|A‖A|nt	2.44	2.23	-2.02	1.06	-2.36	0.81
	/B@PFL (Indiv.)	/B@PFL, n.t.	/B|B|N|I‖A|nt	2.82	1.95	-1.47	0.31	2.67	0.23
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.24	0.95	0.04	0.84	0.98	0.79
Point charge	/B@PFL (All)	/B@PFL, n.t.	/B|B|P|A‖A|nt	1.89	1.63	-1.38	0.83	-0.35	0.77
	/B@PFL (Indiv.)	/B@PFL, n.t.	/B|B|P|I‖A|nt	2.69	2.20	-2.00	0.69	-0.11	0.58
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	3.03	2.49	-2.40	0.51	0.56	0.53
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

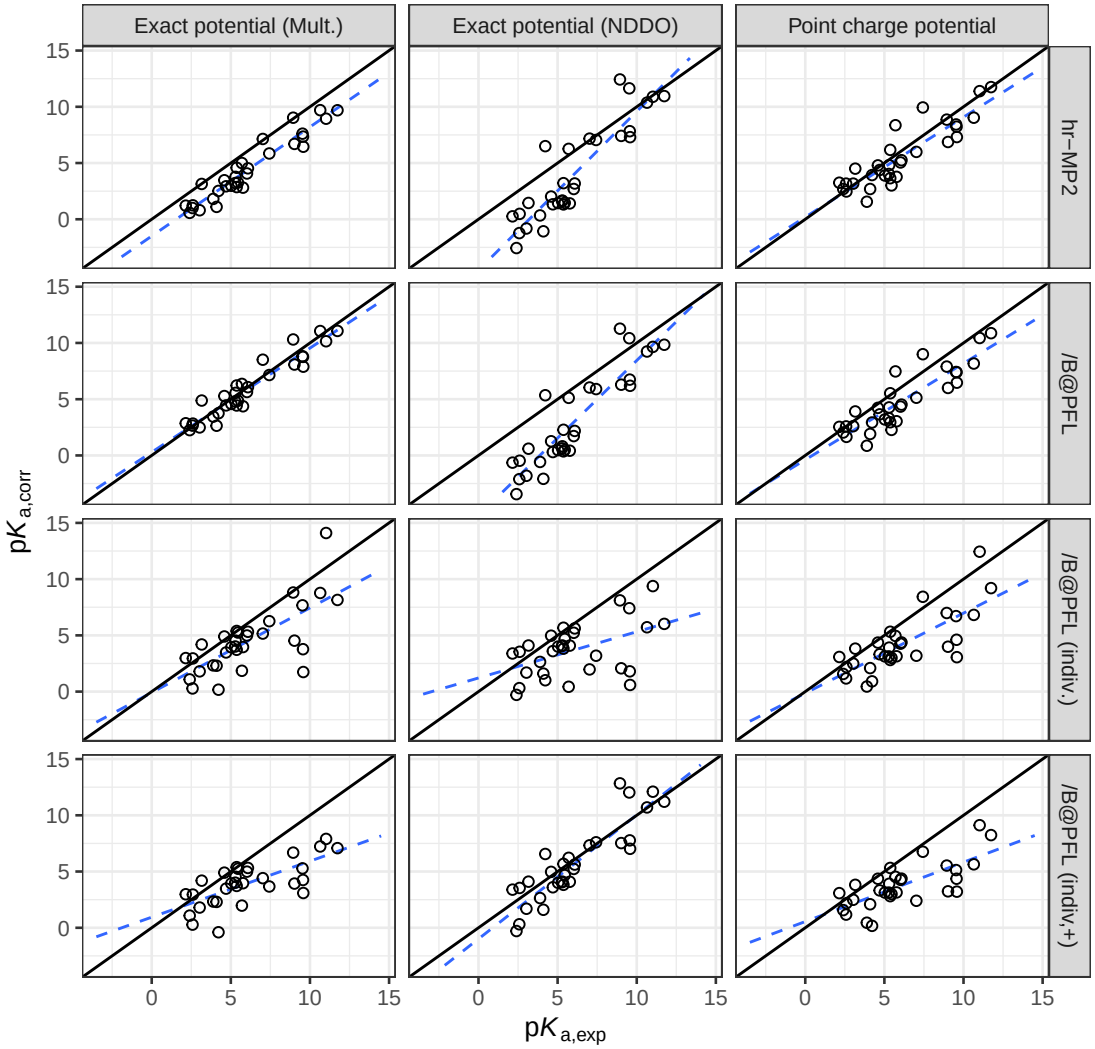

Figure 5.12.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory and the *hr*-MP2 pK_a correction parameters.[24] The rows show the respective PMV correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text.

necessary to apply an individual correction to charged and uncharged macrostates. Since all these NDDO models have been evaluated with the *hr*-MP2 pK_a parameters, it can also be assumed that only the application of this modified individual PMV correction is able to extrapolate to the free energy difference of the *hr*-MP2 level of theory, which is used as the functional argument of the pK_a correction model.

However, when this correction scheme is applied to the remaining two ESP approximations, a different outcome is obtained. The point charge based ESP gives the best result when the *hr*-MP2 parameters are applied to both the PMV and the pK_a correction (*hr*|B|P|A‖A), giving an RMSE of 1.38 and an MSE of −0.48, close to the reference model, although this model was evaluated using a more approximate ESP.

Replacing the *hr*-MP2 with the /B@PFL PMV correction (/B|B|P|A‖*hr*|A) increases the RMSE to a value of 1.81 and decreases the MSE to −1.25, again indicating that this model underestimates the experimental values.

The application of the individual PMV correction at the /B@PFL level (/B|B|P|I‖*hr*|A) further worsens the results. Here a high RMSE of 2.62 is obtained. Furthermore, in contrast to the NDDO models, changing the parameter set of the anionic species to that of the cationic species (/B|B|P|I+‖*hr*|A) does not improve the result, but leads to an even worse RMSE of 2.95. From figure 5.12 it is clear that the high RMSE values are mainly caused by the under-prediction of a cluster of experimental values around 10.

However, the best models based on the *hr*-MP2 pK_a correction can be obtained with the multipole based ESP. The *hr*-MP2 PMV correction (*hr*|B|E|A‖A) gives predictions with an RMSE of 1.86 and a high coefficient of determination of 0.91. Although the predicted values are only slightly scattered around the descriptive regression line and a good slope parameter m' of 0.97 is obtained, the negative intercept parameter b' and the negative MSE, as well as figure 5.12, again indicate that the experimental values are underestimated.

This problem can be corrected by applying the /B@PFL correction with a single parameter set (/B|B|M|A‖*hr*|A). This gives similarly good values for the parameters R^2 and m' as before, but compensates for the underestimation of the experimental pK_a values. This results in an excellent RMSE value of 0.85. This model is therefore able to outperform the prediction quality of the *hr*-reference model for the SAMPL6 dataset, which achieves an RMSE of 1.13. As will be shown later in this chapter, this effect can be attributed to the c_q parameter.

As before for the point charge based model, the application of the two individual corrections does not lead to any further improvement of the predictions and to a higher variation of the predicted pK_a values.

It is noticeable that the best NDDO model, like the *hr* reference model, gives a slope parameter m' greater than 1, while the best models with the multipole and the point charge based approximation of the ESP give an m' of less than 1. This is in agreement with the previous results of the parameterisation data sets, where the latter also gave similar results.

Although accurate, none of the previous best performing models can be considered consistent, as the applied *hr*-MP2 pK_a was parameterised for the *hr*-MP2 PMV correction and subsequently the corresponding *high* level of theory. The purpose of applying this inconsistent parameterisation was to test whether the newly parameterised /B@PFL PMV corrections are able to extrapolate to the free energy difference of the *hr* reference method. In the following, the *hr*-MP2 pK_a correction is replaced by the /B@PFL pK_a corrections, resulting in a more consistent parameterisation. The evaluation of the *hr*-MP2 PMV correction is omitted, as no pK_a correction was parameterised at this level. The corresponding statistical quantities can be found in table 5.15. The predicted pK_a values are shown in figure 5.13.

Regarding the dependence of the pK_a predictions on the ESP approximations used, a similar picture emerges as before. Again, the results of the multipole and point-charge based potentials are similar, while the NDDO-ESP based models show different tendencies. Furthermore, it can be seen that the PMV corrections, which already gave good predictions before, also give good results with the new pK_a correction parameters.

This effect is particularly evident for the PMV correction using a single parameter set, /B@PFL and the multipole based potential (/B|B|M|A‖A). A very good RMSE of 1.13 can be achieved here, resulting in almost identical prediction quality on this test data set compared to the *hr* reference, as indicated by the same RMSE value. However, the ONIOM model has a significantly lower slope parameter of 0.73 compared to the *hr* reference of 1.17. This lower value results from an underestimation of the experimental values above about 8.0 pK_a units.

As with the previous models, the application of the individual corrections "/B@PFL (Indiv.)" and "/B@PFL (Indiv.,+)" (/B|B|M|I‖A and /B|B|M|I+‖A) leads to a deterioration in the prediction quality. Only RMSE values of 2.58 and 2.91 and low coefficients of determination of 0.52 and 0.50 are achieved.

This behaviour can also be observed for the point charge based models. An RMSE of 1.62 is obtained for the "/B@PFL (All)" correction, while this value deteriorates to 2.35 and 2.68 for the two individual corrections.

For the NDDO potential, the PMV corrections which previously gave good results also give the best results here. The PMV corrections with a single parameter set "/B@PFL (All)" (/B|B|N|A‖A) and three separate parameter sets for the respective charge states "/B@PFL (Indiv.)" (/B|B|N|I‖A) each produce predictions with reasonable accuracy and RMSE values of 2.28 and 2.72 respectively. Only the modified PMV correction "/B@PFL (Indiv.,+)" (/B|B|N|I+‖A) returns a satisfactory result with an RMSE of 1.26, only slightly worse than the *hr* reference model.

During the parameterisation of the pK_a corrections in section 5.2 it became clear that the prediction quality of some models on the training dataset was influenced by the substance class "thiols", so additional pK_a corrections were parameterised without this class. This pK_a correction was also tested on the SAMPL6 dataset in combination with the previous PMV corrections. The corresponding results are shown in table 5.15 and

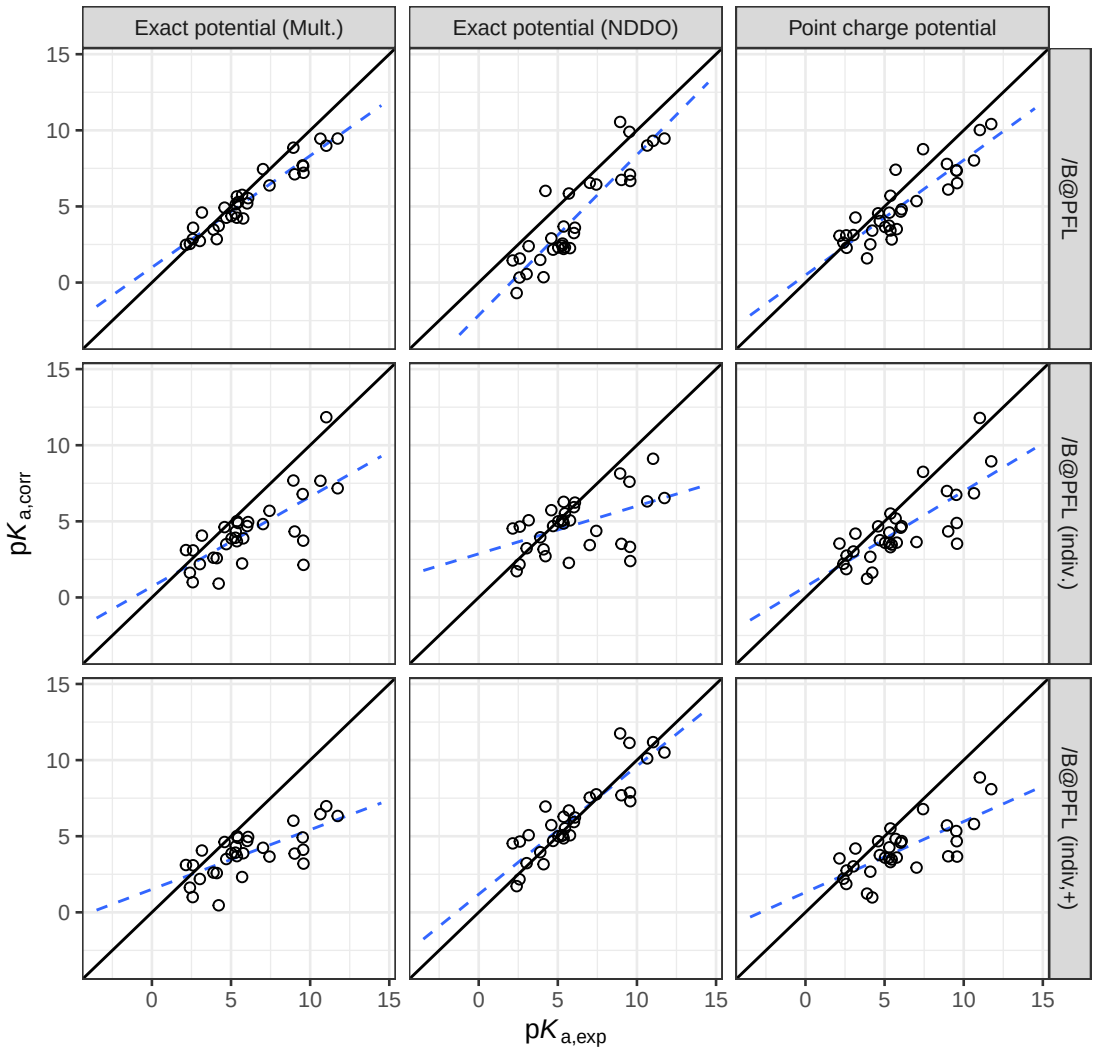

Figure 5.13.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory and the ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11. The rows show the respective PMV correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text.

figure 5.14.

The use of this alternative pK_a correction has the consequence for all multipole models that the prediction quality, as measured by the RMSE, decreases. The "/B@PFL (All)" model (/B|B|M|A‖A|nt), which previously gave an RMSE of 1.13 and thus an identical result to the reference model, now gives a worse RMSE of 1.39. The results of the two individual corrections also continue to deteriorate. However, due to the consistently poor results of these models for all three pK_a corrections, it can be assumed that the individual PMV correction models are not suitable for this potential and should be discarded.

An identical trend can be seen again for the point charge models. The PMV correction model with a single parameter set for all charge states (/B|B|P|A‖A|nt) also gives worse results than before, with an RMSE of 1.89. The same effect as before can also be observed for the individual corrections, so that all models based on them should also be discarded.

Only the NDDO potential shows a good prediction quality for the modified individual PMV correction "/B@PFL (Indiv.,+)" (/B|B|N|I+‖A|nt) with an RMSE of 1.24. However, this is only a slight improvement over the model parameterised on the full Klicić data set (/B|B|N|I+‖A), which gave an almost identical RMSE of 1.26. Applying the thiol-free correction to the remaining two PMV corrections leads to a further deterioration in the results. Again, all previous models resulting from these two PMV corrections did not allow reliable predictions of the experimental pK_a values. Accordingly, these models should also be discarded.

Based on the results for all three pK_a corrections, it can be preliminarily concluded that both the *hr* and the "/B@PFL (All)" correction lead to models with good prediction quality for the multipole-based and point-charge ESP. Particularly noteworthy is the model resulting from the combination of the "/B@PFL (All)" PMV-correction in combination with the pK_a correction and the multipole ESP. An excellent RMSE of 0.85 can be obtained.

For the NDDO potential, only the *hr*-PMV correction and modified individual PMV correction "/B@PFL (Indiv.,+)" lead to results with a high coefficient of determination of the descriptive regression, while only the last correction gives results comparable to the prediction quality of the *hr*reference model.

Overall, it can be observed that the PMV correction has a greater impact on the quality of the resulting prediction model than the associated pK_a correction. For example, with the "/B@PFL (All)" PMV correction and the *hr*-pK_a correction, the best prediction model (/B|B|M|A‖*hr*|A) achieves an RSME of 0.85. The application of the two reparameterised /B@PFL-pK_a corrections (/B|B|M|A‖A and /B|B|M|A‖A|nt) also leads to good predictions with RMSE values of 1.13 and 1.39. In all three cases, models with a high R^2 of the descriptive regression are obtained, with the predicted pK_a values scattering only slightly around these values. On the other hand, there is no PMV correction that produces a result with low prediction quality and low R^2 for a given pK_a correction that can be significantly improved by applying a different pK_a correction.

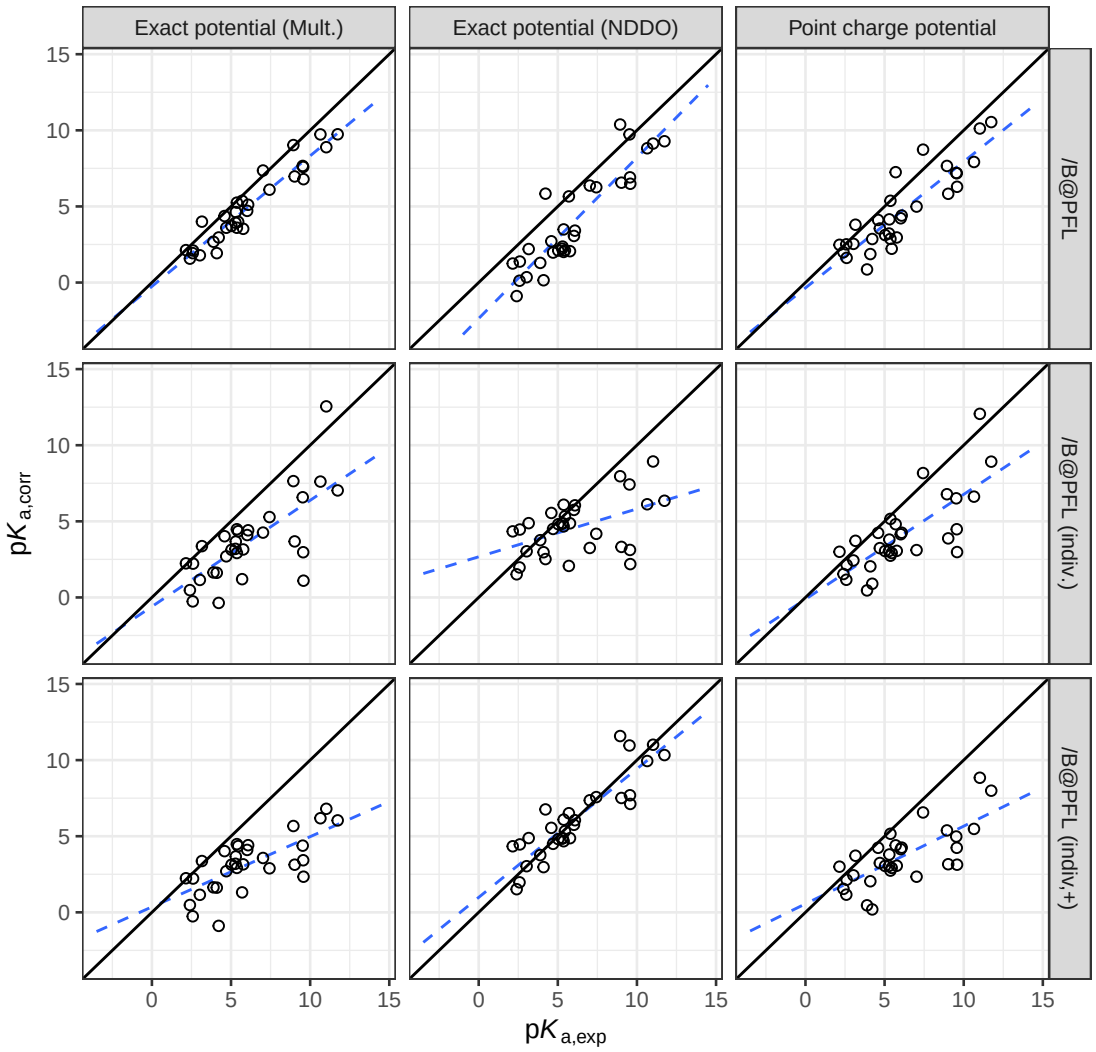

Figure 5.14.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory and the thiol-free ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.13. The rows show the respective PMV correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text.

This is to be expected, since the linear two-parameter model of the pK_a correction simply scales the free energy difference of the macrostates and adds a constant. However, if the differences used as input for the pK_a correction already have a large scatter, this cannot be corrected by this simple linear correction model.

Properties of the charge correction parameter

So far, a number of prediction models have been presented using the newly parameterised /B@PFL corrections, which were able to give results close to those of the *hr* reference model. These models can be considered consistent in their parameterisation, as both the PMV and pK_a corrections have been parameterised at the same theoretical level. It is generally expected that this consistent parameterisation approach should yield the best possible prediction quality. However, the best-performing model so far on the SAMPL6 dataset is composed of a /B@PFL-PMV correction and a *hr*-pK_a correction and must therefore be considered inconsistent. In order to explain this rather surprising result, it is necessary to examine the role played by the charge parameter c_q of the PMV correction in the context of the pK_a prediction.

The PMV correction used in this work uses two parameters c_V and c_q. For the pK_a problem discussed here and a conformer c in the microstate t of the macrostate i they give rise to the free energy contributions $c_V V_{\mathrm{m},itc}$ and $c_q q_{\mathrm{sol},i}$ respectively. The former naturally shows a dependence on the geometry of the molecule and is expected to differ for each conformer c, hence the indices itc on the partial molar volume, while the solute charge and therefore the second free energy contribution is identical for each microstate within a given macrostate. For the purpose of the following discussion, the PMV correction is therefore rewritten as

$$G^{\mathrm{corr}}_{\mathrm{sol},itc} = G'_{\mathrm{sol},itc} + c_q q_{\mathrm{sol},i}, \tag{5.10}$$

separating the charge correction term from the remaining free energy contributions, denoted $G'_{\mathrm{sol},itc}$. Here and in the following the prime symbol indicates that the charge term has been excluded.

This independence of the microstate charge term also means that the partition function, which defines the free energy of the microstate can be split into the product

$$G_{it} = -RT \ln \left[\sum_c \exp\left(-\beta \left(G'_{itc} + c_q q_{\mathrm{sol},i}\right)\right) \right] \tag{5.11}$$

$$= -RT \ln \left[Z_{q,i} Z'_{it} \right], \tag{5.12}$$

where

$$Z_{q,i} = \exp\left(-\beta c_q q_{\mathrm{sol},i}\right) \tag{5.13}$$

and

$$Z'_{it} = \sum_c \exp\left(-\beta G'_{itc}\right). \tag{5.14}$$

This can be simplified to

$$G_{it} = c_q q_{\mathrm{sol},i} - RT \ln Z'_{it}. \tag{5.15}$$

Similarly, the macroscopic free energy can be split into the charge correction term and the remaining contributions

$$G_i = -RT \ln \left[\sum_t \exp\left(-\beta c_q q_{\mathrm{sol},i} + \ln Z'_{it}\right) \right] \tag{5.16}$$

$$= -RT \ln \left[Z_{q_i} Z'_i \right] \tag{5.17}$$

$$= c_q q_{\mathrm{sol},i} - RT \ln Z'_i, \tag{5.18}$$

where

$$Z'_i = \sum_t Z'_{it}. \tag{5.19}$$

This expression then has to be inserted into the linear equation used for the pK_a calculation, already presented in equation 5.7. For the purposes of this discussion, this pK_a correction is rewritten as

$$\mathrm{p}K_{\mathrm{a,corr},i} = M \left(G_{i+1} - G_i\right) + b, \tag{5.20}$$

using $M = m/(RT \ln 10)$. The simplification of this expression gives

$$\mathrm{p}K_{\mathrm{a,corr},i} = MRT \ln \frac{Z'_i}{Z'_{i+1}} - Mc_q e + b. \tag{5.21}$$

By identifying the first additive term as the difference of the macroscopic free energies excluding the charge correction contribution, this equation can also be written as

$$\mathrm{p}K_{\mathrm{a,corr},i} = M \left(G'_{i+1} - G'_i\right) - Mc_q e + b. \tag{5.22}$$

Since the definition of the pK_a as a deprotonation reaction means that the charge difference between the macrostates is always equal to an elementary charge, the charge correction contribution to the pK_a becomes $-Mc_q e$.

Unlike the first additive term of this equation, which depends on the two macrosopic partition functions and contains the parameter c_V, the charge correction contribution is constant for any pK_a value and given parameterisation. It therefore plays a similar role to the parameter b of the pK_a correction, simply shifting the predicted pK_a by a constant value.

This finding can now be used to qualitatively rationalise some of the above findings. First, the shift between *hr* and /B@PFL PMV correction is examined using the *hr*-pK_a correction. Here the /B@PFL PMV correction shows a clear shift towards larger pK_a values. This can be clearly seen in figure 5.12 and by comparison of the respective MSE

and b'. Since the same pK_a correction was used for both models, the m and b parameters are identical. The c_V parameters of the two PMV models are also almost identical. Only the c_q parameter is significantly different between the two. Based on this observation and the previously described influence of c_q, it can be assumed that this shift is therefore caused by this charge parameter, while the underestimation of the pK_a values of the *hr* parameterisation is corrected by applying the /B@PFL correction. Contrary to the consistent *hr* parameterisation, the other model using the /B@PFL PMV parameters must be considered inconsistent and the parameters m and c_q in the additive charge term must be considered independent, as the *hr*-pK_a correction was parameterised using a different set of PMV correction parameter, than those that are applied here. It can therefore be assumed that the exact correction of the underestimation of the pK_a values and thus the high prediction quality of the model from the /B@PFL PMV correction and the *hr*-pK_a parameters is a fortunate cancellation of errors. The charge term takes exactly the value needed here to compensate for the underestimation of the pK_a values.

ONIOM-PCM/X geometries

Based on the insights gained from the first pK_a predictions on the B3LYP-PCM-optimised structures, and in particular the observed deviation from experiment and the coefficient of determination R^2, a subset of models has been selected to test their prediction quality at more approximate levels of optimisation. The selected models and corresponding results obtained on the ONIOM-PCM/X optimised structures are shown in table 5.16 and figure 5.15. The corresponding predicted pK_a values are given in tables 25 to 33 in the appendix.

Compared to the previous B3LYP-PCM geometries, these reoptimised structures generally show slightly worse results. The RMSE obtained from the previous best model based on the "/B@PFL (All)" PMV correction, the *hr*-pK_a correction and the multipole ESP (/B|B|M|A‖*hr*|A) worsens slightly from a value of 0.85 to 0.95. Similarly, the second best model (/B|B|M|A‖A), which gave an RMSE of 1.13, gives a value of 1.27 for the ONIOM geometries. All other multipole-based models also show a slight decrease in prediction quality. In addition, a decrease in R^2 can be observed, further indicating that the predicted pK_a values are more widely distributed around the regression line.

An identical behaviour can be observed for the point charge approximation of the ESP. All models show a slight increase in RMSE. As an example, the best point charge model using both *hr* parameterisations (*hr*|B|P|A‖A) now gives an RMSE of 1.52, while for the B3LYP geometries a slightly better value of 1.38 could be obtained.

Again, the trend for the NDDO potential models is slightly different from the other two ESP approximations. For both prediction models using the single parameter set PMV correction "/B@PFL (All)" an increase in RMSE is observed. On the other hand, all three models using the modified individual PMV correction "/B@PFL (Indiv.,+)" seem to be unaffected by the reduction in the optimisation level and give almost identical

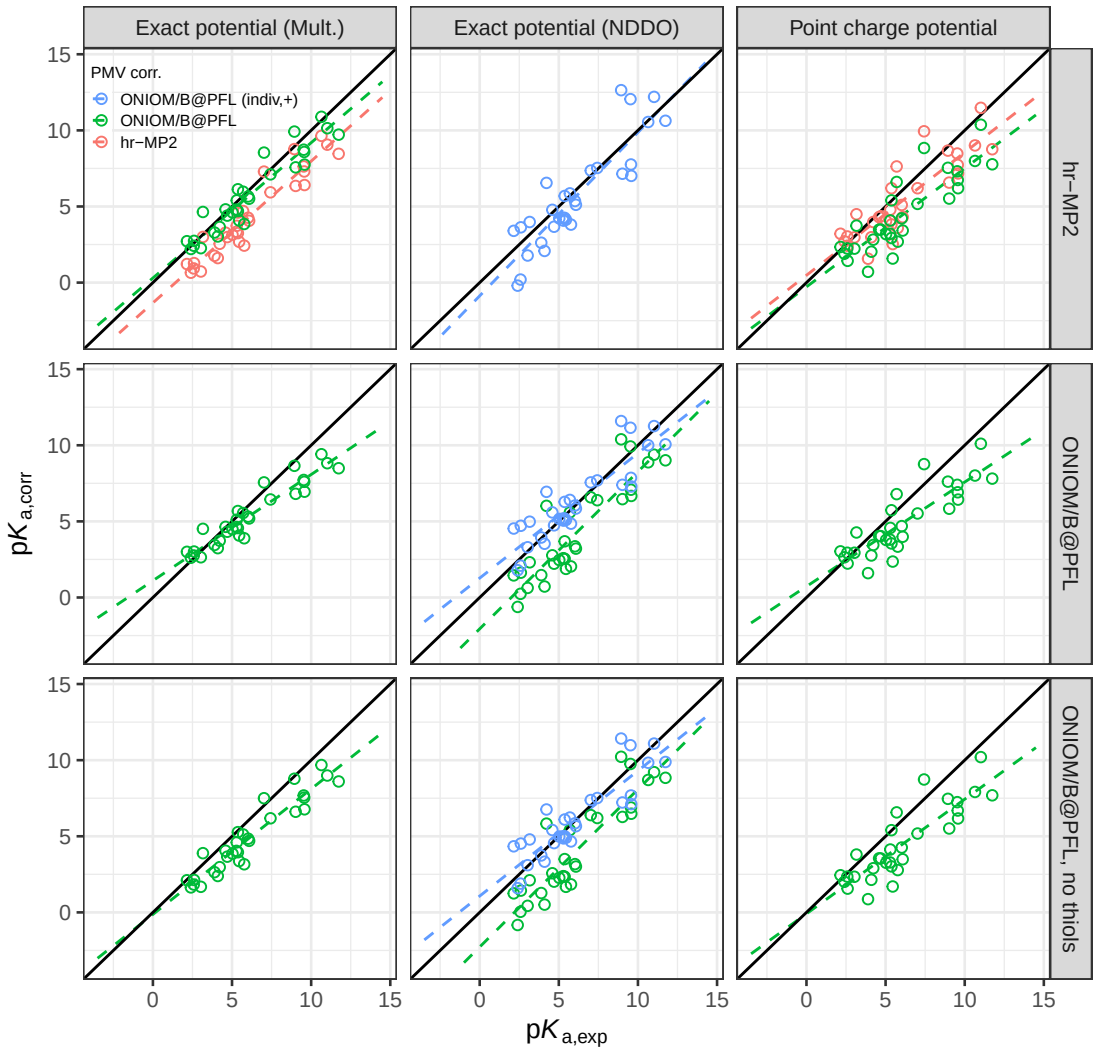

Figure 5.15.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B level of theory on PCM conformers reoptimised at the ONIOM(B3LYP/6-311+G**:PM6)-PCM/X level of theory. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv.,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

Table 5.16.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B single point calculations on ONIOM-PCM/X optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.5, and pK_a correction parameters can be found in table 5.11 or table 5.13 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.95	1.77	-1.75	0.93	-1.33	0.90
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	0.95	0.77	-0.37	0.89	0.30	0.89
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.27	0.98	-0.72	0.70	1.11	0.89
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	1.49	1.27	-1.20	0.82	-0.13	0.89
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|B|N|A‖A	2.31	2.10	-1.86	1.03	-2.06	0.81
	/B@PFL (All)	/B@PFL, n.t.	/B|B|N|A‖A|nt	2.47	2.26	-2.06	1.04	-2.27	0.81
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.58	1.32	-0.42	1.07	-0.87	0.79
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.26	0.94	0.19	0.82	1.27	0.79
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.25	0.96	0.00	0.82	1.07	0.79
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	1.52	1.26	-0.68	0.81	0.48	0.75
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	2.12	1.82	-1.62	0.78	-0.28	0.74
	/B@PFL (All)	/B@PFL	/B|B|P|A‖A	1.81	1.52	-1.17	0.69	0.73	0.74
	/B@PFL (All)	/B@PFL, n.t.	/B|B|P|A‖A|nt	2.10	1.79	-1.58	0.75	-0.09	0.74
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

results. Here, the best model using this modified correction and the thiol-free pK_a correction (/B|B|N|I+‖A|nt) gives an RMSE of 1.25. In contrast, the previous set of geometries gave an almost identical RMSE of 1.24.

PM6-PCM geometries

In order to assess whether an even more cost-effective and approximate level of optimisation than the ONIOM-PCM/X level of theory could be applied to give acceptable pK_a predictions, additional reoptimisations were performed at the *lr* level of theory, i.e. PM6-PCM. The results are given in table 5.17 and figure 5.16. The corresponding pK_a values are given in tables 34 to 42 in the appendix.

Here the effects of the reduced optimisation level are more drastic. Again starting with the best performing model on the B3LYP-PCM geometries, i.e. the multipole ESP model with the "/B@PFL (All)" PMV correction in combination with the *hr*-pK_a correction (/B|P|M|A‖*hr*|A), it can be seen that significantly worse results are obtained for the new set of PM6-PCM geometries. An RMSE of 1.95 is obtained, compared to the previous excellent result of 0.85. A similar deterioration in prediction quality is observed for

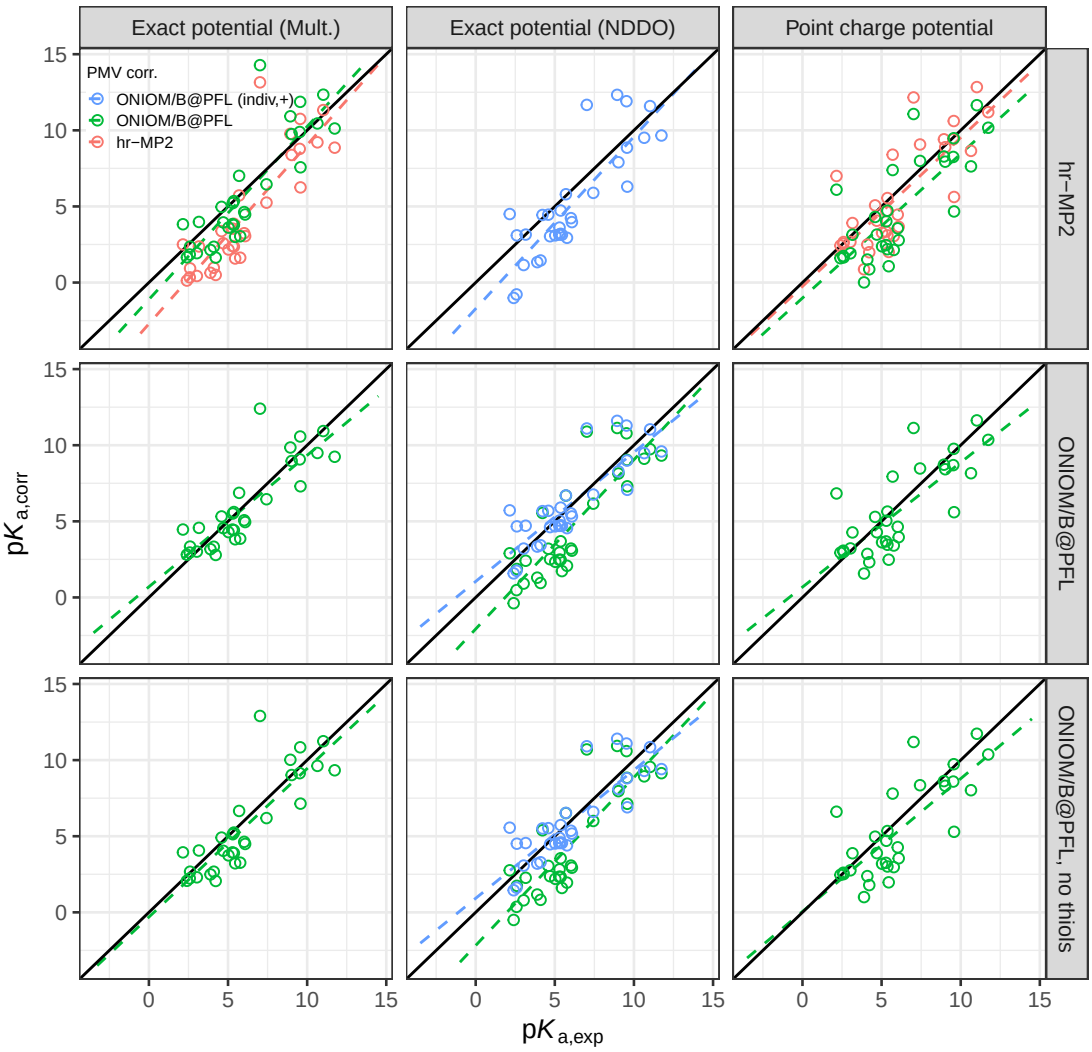

Figure 5.16.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B level of theory on PCM conformers reoptimised at the PM6-PCM level of theory. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv.,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

Table 5.17.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B single point calculations on PM6-PCM-optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.4, and pK_a correction parameters can be found in table 5.10 or table 5.12 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	2.59	2.22	-1.65	1.18	-2.71	0.73
	/B@PFL (All)	*hr*-MP2	/B|P|M|A‖*hr*|A	1.95	1.46	-0.29	1.14	-1.11	0.72
	/B@PFL (All)	/B@PFL	/B|P|M|A‖A	1.49	1.09	-0.12	0.86	0.70	0.72
	/B@PFL (All)	/B@PFL, n.t.	/B|P|M|A‖A|nt	1.69	1.26	-0.46	0.98	-0.31	0.72
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|P|N|A‖A	2.27	2.07	-1.40	1.11	-2.07	0.74
	/B@PFL (All)	/B@PFL, n.t.	/B|P|N|A‖A|nt	2.35	2.15	-1.55	1.10	-2.18	0.74
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|P|N|I+‖*hr*|A	2.17	1.85	-0.93	1.13	-1.72	0.71
	/B@PFL (Indiv.,+)	/B@PFL	/B|P|N|I+‖A	1.51	1.15	0.13	0.85	1.05	0.71
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|P|N|I+‖A|nt	1.50	1.19	-0.03	0.84	0.92	0.71
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	2.14	1.63	-0.39	0.98	-0.25	0.61
	/B@PFL (All)	*hr*-MP2	/B|P|P|A‖*hr*|A	2.48	2.04	-1.34	0.95	-1.02	0.60
	/B@PFL (All)	/B@PFL	/B|P|P|A‖A	1.91	1.48	-0.41	0.82	0.68	0.60
	/B@PFL (All)	/B@PFL, n.t.	/B|P|P|A‖A|nt	2.07	1.60	-0.72	0.87	0.05	0.60
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

the analogous model using both *hr* correction models (*hr*|B|E|A‖A), where an RMSE of 2.59 is obtained compared to the previous value of 1.86. In addition, the coefficient of determination decreases for both models. The drastic deterioration in prediction quality is mainly due to a significant overestimation of the experimental pK_a of 7.02 of molecule SM03. Here, the aforementioned models using the *hr*-pK_a or the /B@PFL PMV correction predict pK_a values of 13.15 and 14.28 respectively. A similar overestimation of this pK_a can be observed for the remaining two multipole ESP models, although the effect is smaller: 12.40 and 12.90 for the /B@PFL PMV and pK_a corrections parameterised with and without the "thiols" subset, respectively. This may partly contribute to the relatively smaller increase in RMSE for these models compared to the two models that use the *hr*-pK_a correction.

The point charge models also show a similar trend. The two models using the *hr*-pK_a correction show a drastic increase in RMSE, while this increase is smaller for the other two models. The best result for these models on the PM6-PCM geometries is obtained with the "/B@PFL (All)" PMV correction and the /B@PFL pK_a correction (/B|P|P|A‖A), which gives an RMSE of 1.91, which is worse than the value of 1.62 obtained from the calculations on the B3LYP-PCM geometries.

Similarly, the best models based on the modified individual PMV correction "/B@PFL (Indiv.,+)" and the NDDO-ESP (/B|P|N|I+‖A and /B|P|N|I+‖A|nt) show a decrease in their prediction quality to an RMSE of about 1.5 in both cases. For the remaining NDDO models, the effect of the PM6-PCM geometries is smaller, but they show high RMSE values of 2.0 and should therefore be discarded for the ONIOM-EC-RISM/B model.

To summarise the effect of the level of geometry optimisation, it can be concluded that in general there is a decrease in prediction quality as the level of optimisation is reduced. This decrease is more drastic for the PM6-PCM geometries than for the ONIOM-PCM/X optimised structures. This effect will be further evaluated for the less approximate ONIOM-EC-RISM schemes /A and /X.

5.3.5. Evaluating the influence of the ONIOM partitions

While the parameterisation of the pK_a models was performed at the /B@PFL level of theory, i.e. without any ONIOM partitions, the final pK_a predictions on the SAMPL6 test data set were performed on partitioned molecules, as shown in figure 5.10. It is therefore necessary to estimate the effect of the partitions on the pK_a models.

To do this, additional /B@PFL single point calculations were performed on the SAMPL6 dataset. In all cases the calculations were performed on the B3LYP-PCM geometries using the ESP multipole approximation. The pK_a values were obtained from the two previously best performing multipole-based models. Both models use the "/B@PFL (All)" PMV correction from table 5.5 and either the corresponding *hr*-pK_a correction (/B|B|M|A‖*hr*|A) or /B@PFL pK_a correction (/B|B|M|A‖A). See table 5.11 for the corresponding parameters. The statistical results are shown in table 5.18 and figure 5.17. The corresponding pK_a values can be found in tables 44 and 45 in the appendix.

Compared to the corresponding /B results, the /B@PFL models generally result in worse predictions. The *hr*-pK_a correction model at the /B@PFL level gives an RMSE of 1.12, while the calculations on the partitioned molecules give an RMSE of 0.85. However, it should be noted that this model still gives slightly better predictions than the *hr* reference model (RMSE of 1.12 vs. 1.13). Similarly, the RMSE of the model using the /B@PFL pK_a correction model increases from 1.13 to 1.17 when PFL is applied. This model is only slightly worse than the *hr* reference model.

In general, partitioning can be expected to introduce an additional source of error by artificially reducing the system size and approximating the local chemical environment through hydrogen link atoms and EE charges. However, the opposite trend is observed here: The /B@PFL models give worse results compared to the /B models. It is therefore likely that these additional approximations of the chemical environment and their resulting errors lead to a cancellation of the errors of the ONIOM-EC-RISM/B approximation and thus to an improved prediction quality.

By calculating the difference in predicted pK_a values between the /B and correspond-

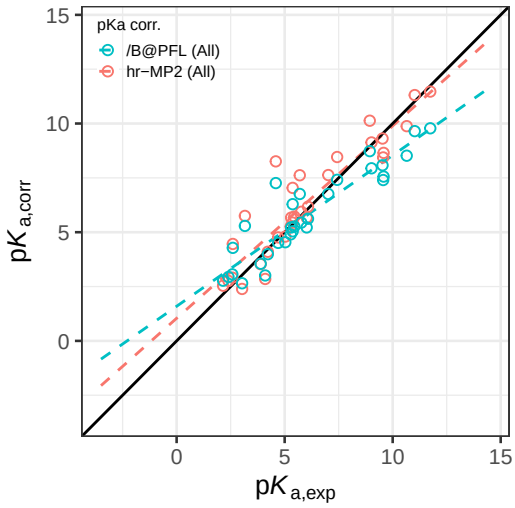

Figure 5.17.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from multipole-ESP based single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory[24] and the corresponding /B@PFL PMV correction from table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"). The colours indicate pK_a corrections based on the "/B@PFL (All)" (/B|B|M|A‖A, blue) and "*hr*-MP2 (All)" (/B|B|M|A‖*hr*|A, red) parameters. See table 5.11 for the corresponding pK_a parameters.

ing /B@PFL models, the effect of the given partitioning can be evaluated directly for each molecule. For the ONIOM-EC-RISM/B models this difference isolates the effect of the ONIOM partitions on the electronic structure calculation. These values are shown in figures 5.18, 5.19 and 5.20 in addition to the deviations from the experiment for the /B model. The first figure shows these pK_a differences for all molecules sharing the scaffolds shown in figure 5.11. The other two figures show these values for the remaining molecules, while the last figure shows the results for SM18, SM23 and SM24 for which multiple partitions were created. The corresponding numerical values can be found in tables 43 to 47 in the appendix.

Single partitions of structures sharing a common scaffold

From figure 5.18 it can be seen that in all cases the *hr*-pK_a correction yields smaller absolute differences between /B@FPL and /B than the /B@PFL correction, while also yielding smaller deviations from the experimental values. For the subsets of molecules that share the first heterocyclic scaffold from figure 5.11, i.e. SM02, SM04, SM07, SM09, SM12 and SM13, the difference between /B and /B@PFL, and thus the partitioning

Table 5.18.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL single point calculations on B3LYP-PCM-optimised geometries. The best performing ONIOM-EC-RISM/B models from the previous sections (see table 5.15 and section 5.3.4 for more context) and the *hr*-reference model using the exact potential from the original SAMPL6 publication[24] are shown in the last three rows. An explanation of the PMV fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. All PMV parameters used for the models in this table can be found in table 5.5, and pK_a correction parameters can be found in table 5.11.

EC-RISM	Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
/B@PFL	Exact (Mult.)	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	1.12	0.75	0.35	0.89	1.04	0.85
		/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.17	0.90	-0.25	0.69	1.59	0.85
/B	Exact (Mult.)	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	0.85	0.71	-0.16	0.93	0.29	0.90
		/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.13	0.89	-0.61	0.73	0.99	0.91
hr-ref.	Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

effect, is smaller than about 0.5 pK_a units in absolute terms in almost all cases.

The largest partitioning effects of -1.0 and -1.57 pK_a units are observed for SM09 and SM13 when the /B@PFL pK_a correction is applied. In addition, the /B model for these two molecules also shows the largest deviations from the experiment, regardless of the pK_a correction applied. For both SM09 and SM13 partitions have been defined that exclude moieties containing methoxy groups from the *model* system. In the case of SM09 this means excluding the anisole moiety, while for SM13 two methoxy groups are excluded. The other molecules of this subset have no such moieties, in fact all of the central fragment substituents only consist of chloride heteroatoms, so it is possible that the exclusion of the ether group is an important factor in the larger deviations observed. However, by visual inspection of all conformers of SM09 it can be confirmed that no direct hydrogen bonds are formed between the nitrogen protonation centres of the central fragment and the ester group, which would potentially require both to be modelled at the *high* level of the theory.

The /B model with the *hr*-pK_a correction gives a pK_a value of 4.44, while the /B@PFL correction gives a value of 4.26. In contrast, the *hr* reference gives a pK_a prediction of 4.17,[24] which, with a given experimental value of 5.37, is similar but slightly worse than the /B models. However, since both the *hr* reference and the ONIOM-EC-RISM models give similar results, it can be assumed that the larger deviations from the experiment are not due to the chosen partitioning of SM09 and hence the ONIOM approximation, but rather to some other effect present in both EC-RISM models.

In contrast to these results, the *hr* reference model predicts a pK_a value of 5.19 for the experimental pK_a of 5.77 of SM13, which is better than the results of the /B models, which are 4.37 and 4.21 respectively. It is therefore more likely than in the case of SM09 that the more drastic splitting of SM13, which cuts off ether groups that are part of the

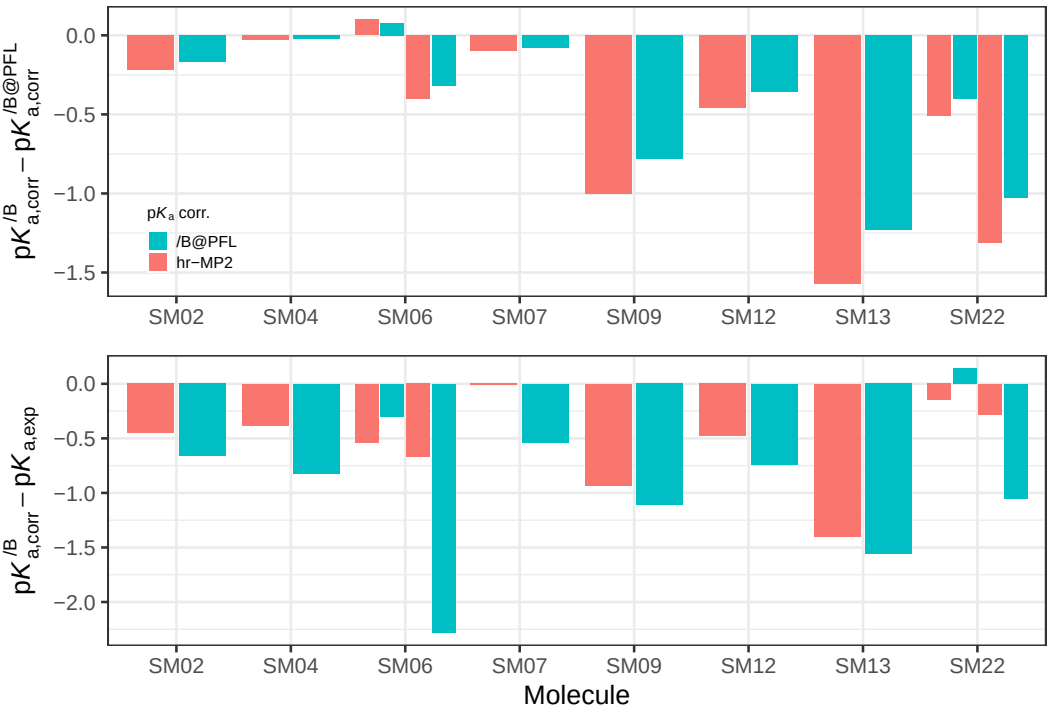

Figure 5.18.: pK_a differences of the molecules sharing the scaffolds shown in figure 5.11. The upper plot shows the difference between the /B and /B@PFL results, while the lower plot shows the deviation from the experiment for the /B model. The results were obtained from calculations on the B3LYP-PCM-optimised structures and corrected with the /B@PFL PMV parameters from table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"). The colours indicate pK_a corrections based on the "/B@PFL (All)" (/B|B|M|A‖A, blue) and "*hr*-MP2 (All)" (/B|B|M|A‖*hr*|A, red) parameters. See table 5.11 for the corresponding pK_a parameters.

delocalised electron system of the central fragment, gives rise to these worse predictions.

For the second subset of SM06 and SM22, with experimental pK_a values of 11.74 and 2.40 and 7.43 respectively, only the prediction of the second pK_a value of SM22 shows a pronounced partitioning effect of more than one pK_a unit. For the remaining pK_a values the effect is less pronounced.

For these two molecules, the /B model with the *hr*-pK_a correction gives accurate predictions compared to the experimental result. In contrast, the /B@PFL pK_a correction generally yields predictions that show larger deviations from the experiment for the high pK_a values of 11.74 and 7.43. This is most likely due to the reduced prediction quality for higher pK_a values for this model, as indicated by the significantly lower value for m'

compared to the model using the pK_a correction.

Single partitions of the remaining structures

Figure 5.19 shows the same analysis as before for the remaining molecules for which only one partitioning was created.

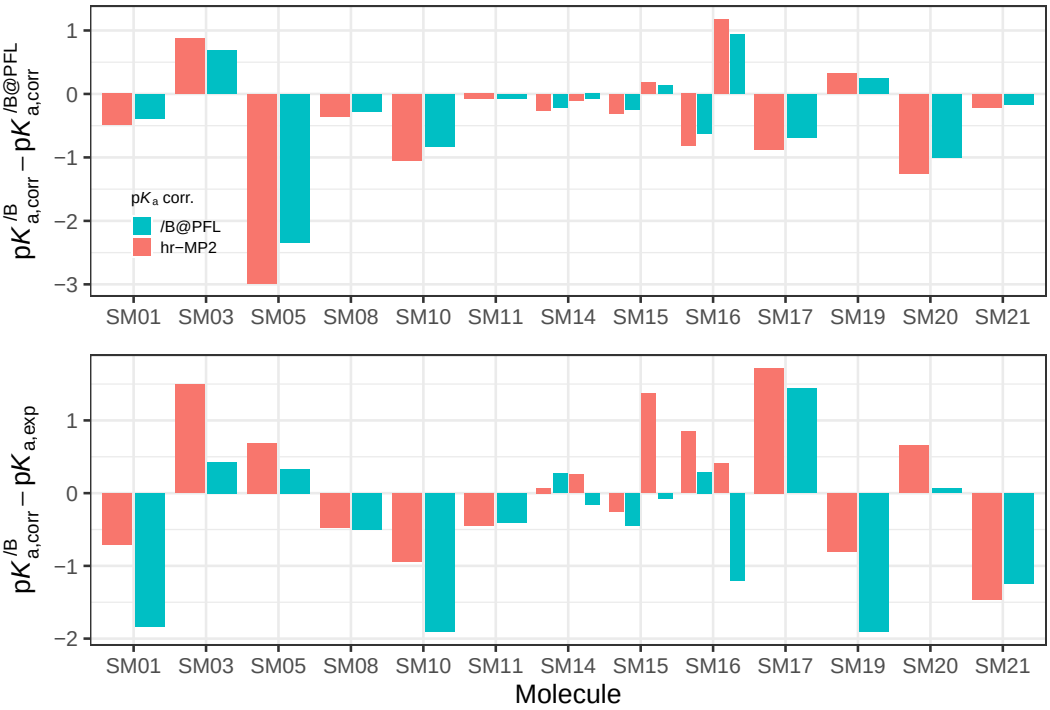

Figure 5.19.: pK_a differences of the single-partition molecules that do not share the scaffolds shown in figure 5.11. The upper plot shows the difference between the /B and /B@PFL results, while the lower plot shows the deviation from the experiment for the /B model. The results were obtained from calculations on the B3LYP-PCM-optimised structures and corrected with the /B@PFL PMV parameters from table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"). The colours indicate pK_a corrections based on the "/B@PFL (All)" (/B|B|M|A‖A, blue) and "*hr*-MP2 (All)" (/B|B|M|A‖*hr*|A, red) parameters. See table 5.11 for the corresponding pK_a parameters.

It can be seen that the absolute partitioning effect is generally smaller than one pK_a unit. This effect is particularly pronounced for SM05, which shows a partitioning effect smaller than -2 pK_a units for both pK_a corrections. At the same time, the deviations from the experimental value for the /B model are smaller than 0.7 pK_a units and thus

significantly better than the *hr* reference model, which gives a deviation of 2.15.[24] Since the /B@PFL model shows a similar overestimation as the *hr* reference, it can be assumed that the partitioning effect counteracts this overestimation and thus leads to accurate pK_a predictions.

Surprisingly, the rather drastic partitioning of SM15, where bonds connecting two fused rings were severed, gives only a small absolute partitioning effect of 0.31 or less, depending on the predicted pK_a value and pK_a correction.

Multiple partitions for SM18, SM23 and SM24

The graphical analysis for SM18, SM23 and SM24, for which four partitions were created, is shown in figure 5.20.

For SM18 the absolute partitioning effect is in most cases less than 1 pK_a unit. Larger effects can be observed for the third pK_a value, where the fourth partitioning P4 gives the largest absolute partitioning effect. Although this partitioning is characterised by more drastic cuts where, similar to SM15, two bonds connecting two fused rings have been severed, it is unlikely that this causes the large partitioning effect for the third pK_a value, since P2 and P3 do not share the cut within the ring system but show similar values of about -1.1 pK_a units. More specifically, the influence of the cut in the ring system can also be evaluated using P3, as this partition is otherwise identical to P4.

While the deviation from the third experimental pK_a value is only slightly larger for P4 compared to all other partitions, the deviation is larger for the other two pK_a values. The best results are obtained for P1 and P3, where the ring system is kept as a single connected fragment in the *model* system. The deviations from the experiment are slightly larger for P2, but still smaller than for P4. It can therefore be assumed that the separation of the conjugated ring system is not ideal and leads to a reduced prediction quality.

In contrast to SM18, only one experimental pK_a value is given for SM23 and SM24. For the former, the smallest partitioning effect is observed for P1, which is to be expected as only the terminal methyl groups were excluded to obtain the *model* system. There is therefore only a minimal difference between the *real* and *model* systems, resulting in a minimal partitioning effect of 0.03 for both pK_a corrections. This partitioning leads to small deviations from the experiment of 0.30 and -0.16 pK_a units for the *hr* and /B@PFL pK_a corrections respectively.

The next smaller *model* system P2 additionally excludes the adjacent methylene group and thus the entire ethyl group of the terminal ester moieties. This partitioning therefore results in a *model* system where this ester is approximated as a carboxylic acid. This drastic approximation may be the cause of the increased partitioning effect of 0.86 and 0.68 for the two pK_a corrections respectively. At the same time the deviation from the experiment increases to values of 0.68 and 1.13.

According to the organisers of the SAMPL6 challenge, one of the oxygen atoms of the

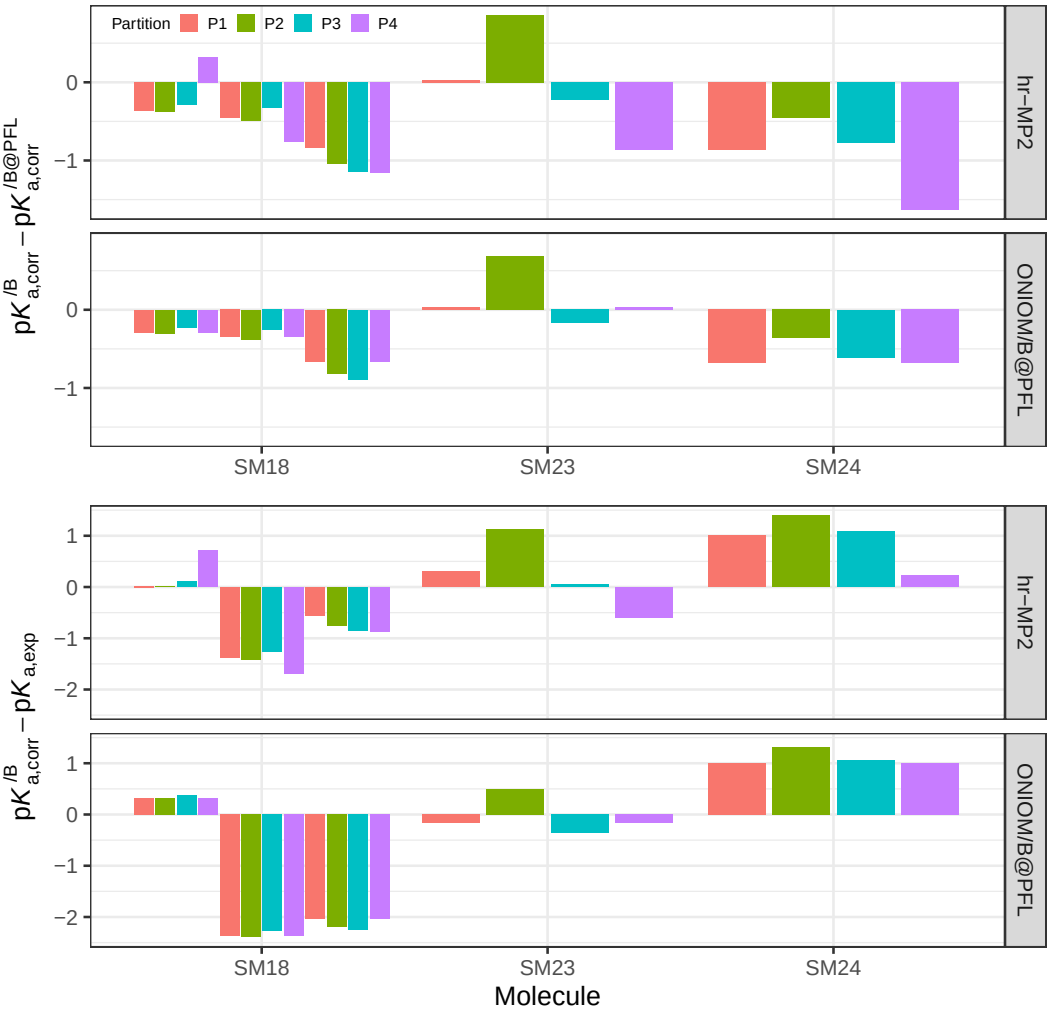

Figure 5.20.: pK_a differences of the molecules with multiple partitions. The upper plots show the difference between the /B and /B@PFL results, while the lower plots show the deviation from the experiment for the /B model. The results are obtained from calculations on the B3LYP-PCM-optimised structures corrected with the /B@PFL PMV parameters from table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"). Subplots differentiate between the "/B@PFL (All)" (/B|B|M|A‖A) and "*hr*-MP2 (All)" (/B|B|M|A‖*hr*|A) pK_a correction parameters. See table 5.11 for the corresponding pK_a parameters. Colours differentiate between P1 (red), P2 (green), P3 (blue) and P4 (purple) partitions.

ester moiety is a protonation centre. As the protonated state of this ester group is not expected to be highly populated in the range of experimentally relevant pH values, it was excluded to create additional smaller partitionings. For P3, the entire ester moiety was excluded from the *model* system. Compared to P2 this reduces the partitioning effect to values of -0.22 for the *hr*-pK_a correction and -0.17 for the corresponding /B@PFL correction. The prediction quality also increases drastically, as measured by the respective deviations from the experiment of 0.05 and -0.36.

P4, the final partitioning for SM23, reduces the *model* system to the core fragment containing all protonation centres. Despite this drastic approximation to the *real* system, this partitioning yields a good prediction quality with deviations from the experiment of -0.59 and -0.16 pK_a units for the respective pK_a corrections.

For SM24 a similar incremental reduction in *model* system size was defined by multiple partitions, i.e. the *model* system size decreases in the order P2, P3, P1, P4. It can be observed that the absolute partitioning effect increases with decreasing *model* system size. For P2 only the two terminal methyl groups of the ether moiety are excluded from the *model* system resulting in a partitioning effect of -0.46. For P3 the entire methoxy group is excluded, giving a partitioning effect of -0.77 pK_a units. The next smaller *model* system P1 is obtained by removing the entire anisole group. This again gives a smaller value of -0.86. Cutting the bonds connecting the two fused rings in P4, similar to SM15 or P4 of SM18, again gives a lower value of -1.16. The absolute partitioning effect thus increases as larger parts of the molecule are excluded from the *model* system. For the second pK_a correction applied here, i.e. the /B@PFL correction, a similar trend can be observed, although the partitions P1 and P4, which yield the smallest values, give the same partitioning effect.

In terms of deviations from the experimental pK_a value, the smallest partition P4 gives the smallest deviation of 0.23 for the *hr*-pK_a correction. In contrast, the *hr* reference model gives a corresponding value of -0.12.[24] The remaining partitions all give results at or above about 1 pK_a unit. Overall, it can be seen that the absolute deviation from the experiment decreases as the size of the *model* system decreases. It is therefore likely that the partitioning error counteracts the errors of the /B approximation.

Based on the results of this section, it can be briefly summarised and concluded that the partitioning of the SAMPL6 molecules improves the prediction quality of the ONIOM-EC-RISM/B model. By calculating the pK_a values at the size extrapolation limit /B@PFL it can be shown that the fundamental approximations of the ONIOM-EC-RISM/B model without any ONIOM partitioning effects give results that are mostly equal in prediction quality to the *hr* reference model, thus validating the model. Since the introduction of the ONIOM partitions into the model, i.e. the transition from the /B@PFL to the /B model, improves the overall prediction quality, it can be concluded that the partitioning effect counteracts some errors of the /B approximation and, in general, of the EC-RISM model. However, it should be emphasised that the /B@PFL model, although giving worse results than the /B model, still yields RMSE values that

are nearly identical to the less approximate *hr*-reference model.

5.3.6. Acidity constant prediction with ONIOM-EC-RISM/A

Now that the prediction quality of the ONIOM-EC-RISM/B model on the SAMPL6 dataset has been extensively discussed and some excellent models have been identified, it is time to focus attention on the other ONIOM-EC-RISM approximations, namely /A and /X. The discussion will start with the evaluation of the former. Since it is expected that the individual ONIOM-EC-RISM schemes will all give similar results due to their methodological proximity, the combinations of PMV and pK_a correction that worked best for the /B scheme are also applied to the /A model.

B3LYP-PCM geometries

The results obtained with the /A model on the B3LYP-PCM geometries are shown in figure 5.21 and table 5.19. The numerical values of the predicted pK_a values are given in tables 48 to 56.

Table 5.19.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A single point calculations on B3LYP-PCM-optimised geometries.[24] The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.5, and pK_a correction parameters can be found in table 5.11 or table 5.13 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.39	1.28	-0.80	1.21	-2.08	0.92
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	1.29	0.92	0.75	1.17	-0.28	0.92
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	0.78	0.59	0.06	0.92	0.56	0.92
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	0.95	0.84	-0.27	1.08	-0.78	0.92
Exact (NDDO)	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.17	0.99	0.52	0.91	1.05	0.85
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.45	1.26	0.91	0.70	2.74	0.85
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.34	1.15	0.72	0.70	2.54	0.85
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	1.33	1.05	-0.81	0.81	0.37	0.85
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	1.94	1.64	-1.59	0.78	-0.23	0.83
	/B@PFL (All)	/B@PFL	/B|B|P|A‖A	1.76	1.40	-1.29	0.68	0.63	0.83
	/B@PFL (All)	/B@PFL, n.t.	/B|B|P|A‖A|nt	2.05	1.75	-1.71	0.75	-0.20	0.83
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

All models obtained from the multipole ESP approximation give an RSME of less than 1.4 pK_a units. From this group of models, the two models using the *hr*-MP2

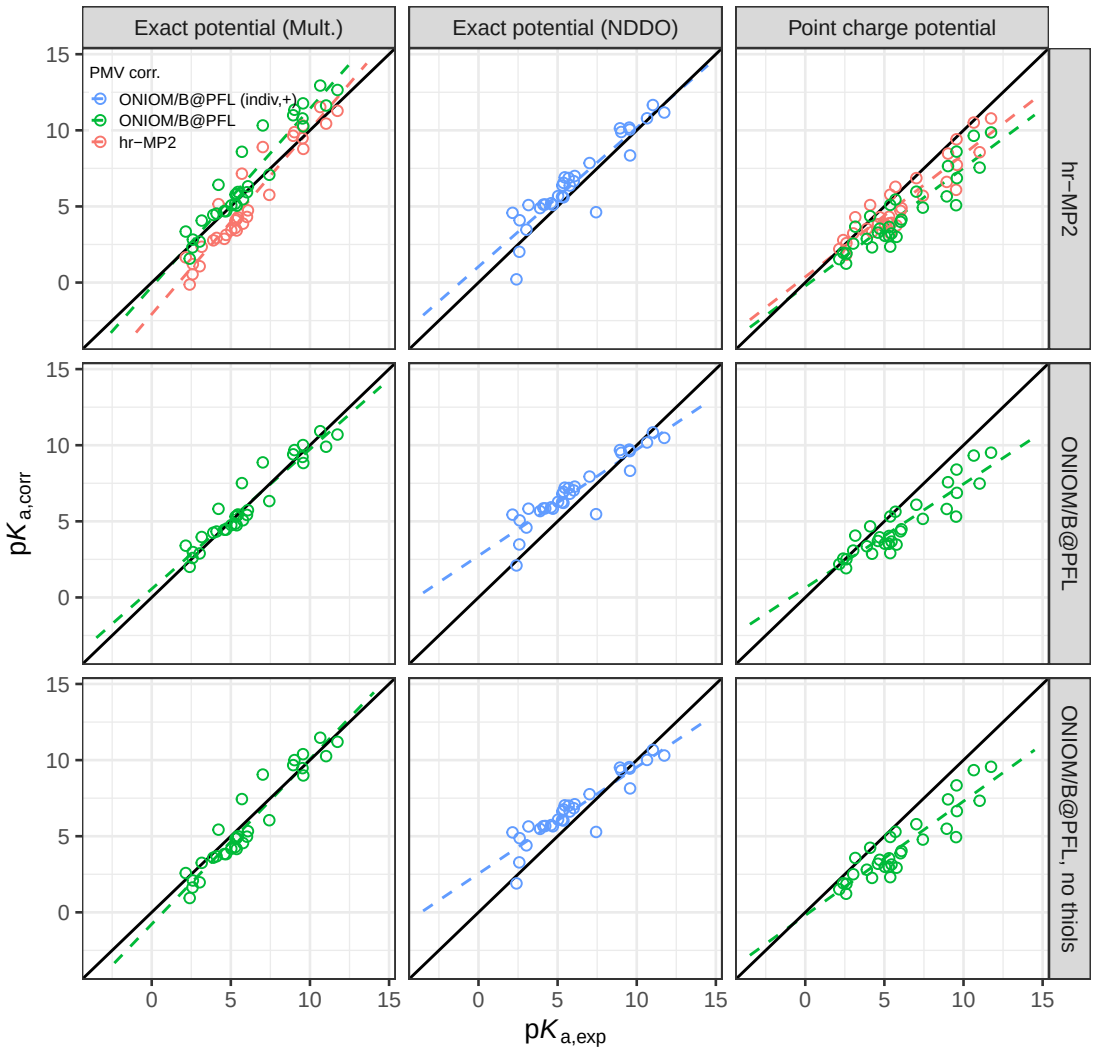

Figure 5.21.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory.[24] The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv.,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

pK_a correction give significantly worse results than those using the /B@PFL models. The combination of the "/B@PFL (All)" PMV correction and the *hr*-pK_a correction (/B|B|M|A‖*hr*|A), which gave the best result for the /B approximation with an RMSE of 0.85, gives an increased RMSE of 1.29 for the /A model. As this model is the only multipole based model that gives an MSE significantly larger than zero and larger pK_a values than the corresponding model using the *hr*-pK_a PMV correction, it can be assumed that the additive contribution of the c_q parameter now overcompensates for the underprediction of the pK_a values and thus reduces the prediction quality. This further suggests that the excellent prediction quality of the corresponding /B model was achieved due to an exact additive compensation of the underprediction of the experimental pK_a.

In contrast, the /B@PFL corrections, which are consistently parameterised in terms of the /B approximation, give excellent RMSE values, outperforming the *hr* reference model on this test data set. In particular, the /A model, which uses the "/B@PFL (All)" PMV correction and the /B@PFL pK_a correction (/B|B|M|A‖A) stand out with a low RMSE of 0.78. The corresponding /A model with the thiol-free pK_a correction also gives a good RMSE of 0.95.

It is noteworthy that for all these models the coefficient of determination always takes a high value of 0.92, indicating that in all cases the predicted pK_a values are only slightly scattered around the regression line.

The selected parameter combinations for the NDDO-ESP result in models that perform significantly worse than the multipole based /A models discussed above, as well as the corresponding NDDO models using the /B approximation. The best model is obtained from the *hr*-pK_a correction (/B|B|N|I+‖*hr*|A) with an RMSE of 1.17. The application of both /B@PFL pK_a corrections increases the RMSE to 1.45 and 1.34 respectively. In all cases the MSE indicates a systematic underprediction of the experimental pK_a values.

For the point charge based models a similar increase in RMSE compared to the /B models is observed, except for the model using both *hr*-pK_a corrections (*hr*|B|P|A‖A). Here an acceptable RMSE of 1.33 is obtained. All other models give significantly worse predictions with an RMSE above 1.76.

For both the NDDO-ESP and the point charge based approximation, a coefficient of determination not greater than 0.85 is obtained, compared to the value of 0.92 for the multipole based models. This indicates a greater dispersion around the regression line.

ONIOM-PCM/X geometries

As for the ONIOM-EC-RISM/B models, the effect of the level of optimisation is investigated by predicting the pK_a values on the reoptimised ONIOM-PCM/X geometries. The corresponding results can be found in figure 5.22 and table 5.20, as well as in tables 57 to 65 in the appendix.

Reducing the optimisation level to the less expensive ONIOM-PCM/X theory results

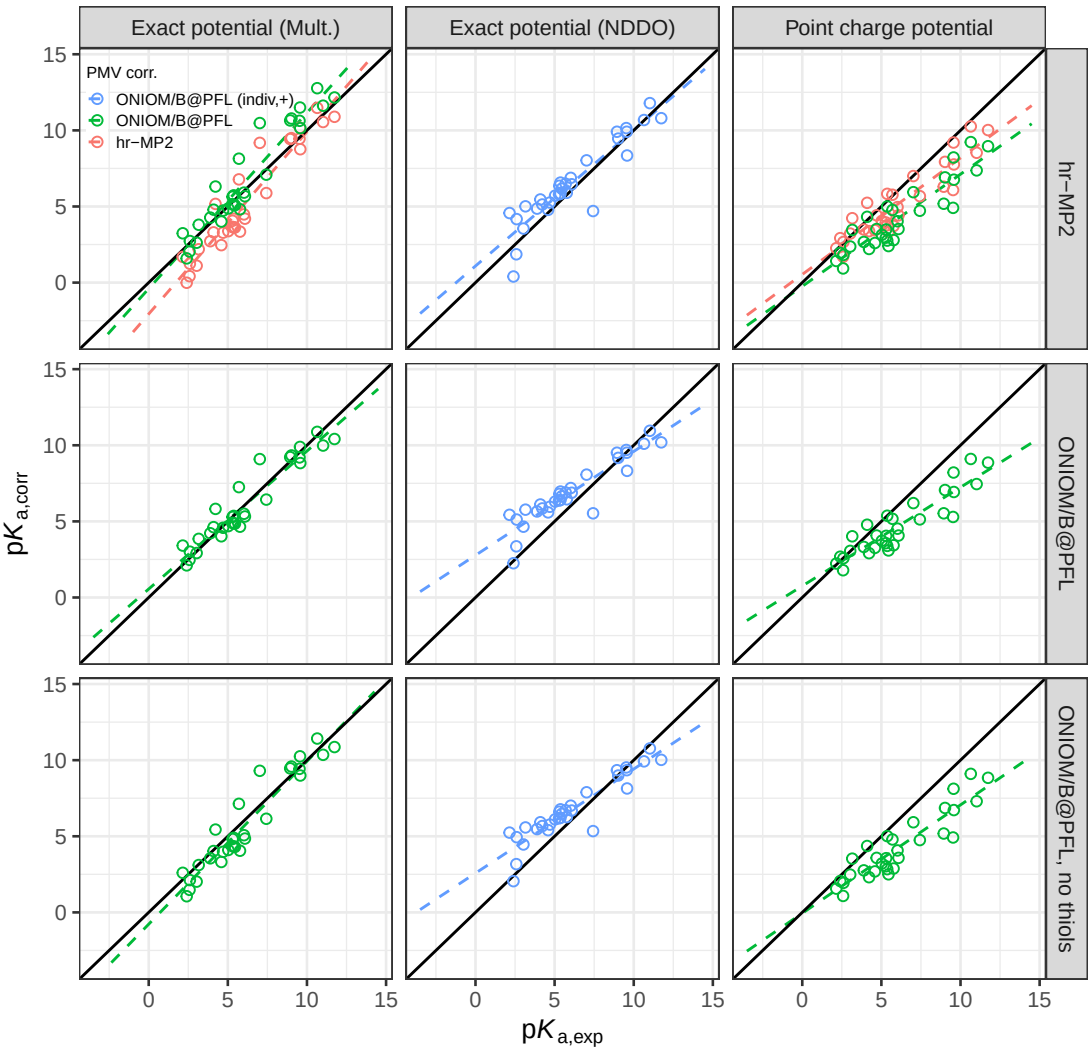

Figure 5.22.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A level of theory on PCM conformers reoptimised at the ONIOM(B3LYP/6-311+G**:PM6)-PCM/X level of theory. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

Table 5.20.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A single point calculations on ONIOM-PCM/X optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.5, and pK_a correction parameters can be found in table 5.11 or table 5.13 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.44	1.30	-0.87	1.20	-2.06	0.91
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	1.19	0.87	0.54	1.16	-0.40	0.91
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	0.81	0.63	-0.01	0.91	0.56	0.91
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	0.99	0.86	-0.35	1.07	-0.77	0.91
Exact (NDDO)	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.12	0.93	0.44	0.89	1.09	0.85
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.42	1.21	0.85	0.68	2.77	0.85
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.32	1.11	0.67	0.68	2.58	0.85
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	1.40	1.12	-0.88	0.76	0.54	0.85
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	2.17	1.87	-1.84	0.74	-0.25	0.83
	/B@PFL (All)	/B@PFL	/B|B|P|A‖A	1.85	1.49	-1.37	0.65	0.76	0.83
	/B@PFL (All)	/B@PFL,n.t.	/B|B|P|A‖A|nt	2.15	1.84	-1.80	0.71	-0.06	0.83
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

in a shift to smaller pK_a values for all multipole ESP models. As these models already showed a systematic underprediction of the experimental pK_a values, as indicated by the negative MSE, the RMSE increases in all cases. For example, the /A model, which includes the "/B@PFL (All)" PMV correction and the /B@PFL pK_a correction (/B|B|M|A‖A) gave an RMSE of 0.78 on the B3LYP-PCM geometries and now gives an increased RMSE of 0.81 on the reoptimised structures. This is of course still an improved result compared to the best performing /B and *hr*-reference models evaluated on the more expensive and less approximate B3LYP-PCM geometries.

The only exception to this trend is the model derived from the "/B@PFL (All)" PMV correction and the *hr*-pK_a correction (/B|B|M|A‖*hr*|A). For the B3LYP-PCM geometries an overprediction could be observed, which is now counteracted by the decrease of the optimisation level. This is the only model where the RMSE improves slightly from 1.29 to 1.19.

A similar trend can be observed for the NDDO-ESP models. Here the MSE decreases for all three models, indicating that there is a systematic shift towards smaller pK_a values due to the application of the ONIOM-PCM/X geometries. At the same time, a slight decrease in the RMSE can be observed. The above argument can also be applied here. The overprediction observed for the B3LYP-PCM structures is compensated by the

application of the lower optimisation level. Therefore, the prediction quality increases in all cases, although this effect is small and none of these models gives a prediction quality close to the excellent results observed for the multipole-based /A models. For example, the application of the *hr*-MP2 pK_a correction (/B|B|N|I+‖*hr*|A) gives an RMSE of 1.12, which is effectively equal to the *hr* reference model, but significantly worse than the aforementioned multipole models.

Also for the point charge based models the same systematic shift towards smaller pK_a values, as measured by the MSE, can be observed as the optimisation level is reduced. Similar to the multipole based models, the previous underprediction is further increased. Thus, the prediction quality continues to decrease. The best model with an RMSE of 1.40 is obtained by applying both *hr* corrections (*hr*|B|P|A‖A). The remaining models give RMSE values above 1.85 pK_a units.

To briefly summarise the results of this section, the application of the ONIOM-PCM/X geometries leads in all cases to a systematic decrease of the predicted pK_a values. As a consequence, the prediction quality of the respective model is increased in comparison to the results of the B3LYP-PCM geometries, if the latter lead to a systematic overprediction, which can be compensated by decreasing the optimisation level. This is the case for all NDDO models and one multipole-based model that uses the "/B@PFL (All)" PMV correction and the *hr*-MP2 pK_a correction. In comparison, all other models show systematically lower pK_a values and therefore show no improvement as the optimisation level decreases.

PM6-PCM geometries

Finally, the same analysis as before can be performed for the PM6-PCM reoptimised structures. Despite the large cost reduction achieved by applying this level of optimisation, the results presented in figure 5.23 and table 5.21 indicate that the resulting prediction quality is generally low and that these structures cannot be used to obtain useful pK_a predictions.

The parameterisation strategy that gave the best multipole-based model with an RMSE of 0.78 on the B3LYP-PCM-optimised structures now gives a significantly worse result of 1.65 (/B|P|M|A‖A). The remaining models also show results with increased RMSE values above 1.83 pK_a units.

The same is true for the other two ESP approximations. The best NDDO model (/B|P|N|I+‖A|nt) gives an RMSE of 1.77, while the best point charge model (/B|P|P|A‖A) gives an RMSE of 1.87. In all cases this is far from the prediction quality obtained by the B3LYP-PCM or the ONIOM-PCM/X optimisations.

In addition, the previous effect of the systematic shift to smaller pK_a values for the ONIOM-PCM/X geometries cannot be observed here. However, a drastic decrease of R^2 is visible for all models. From this and the increase in the MAE it can be concluded that the use of the PM6-PCM conformers are likely to cause a greater scatter of the

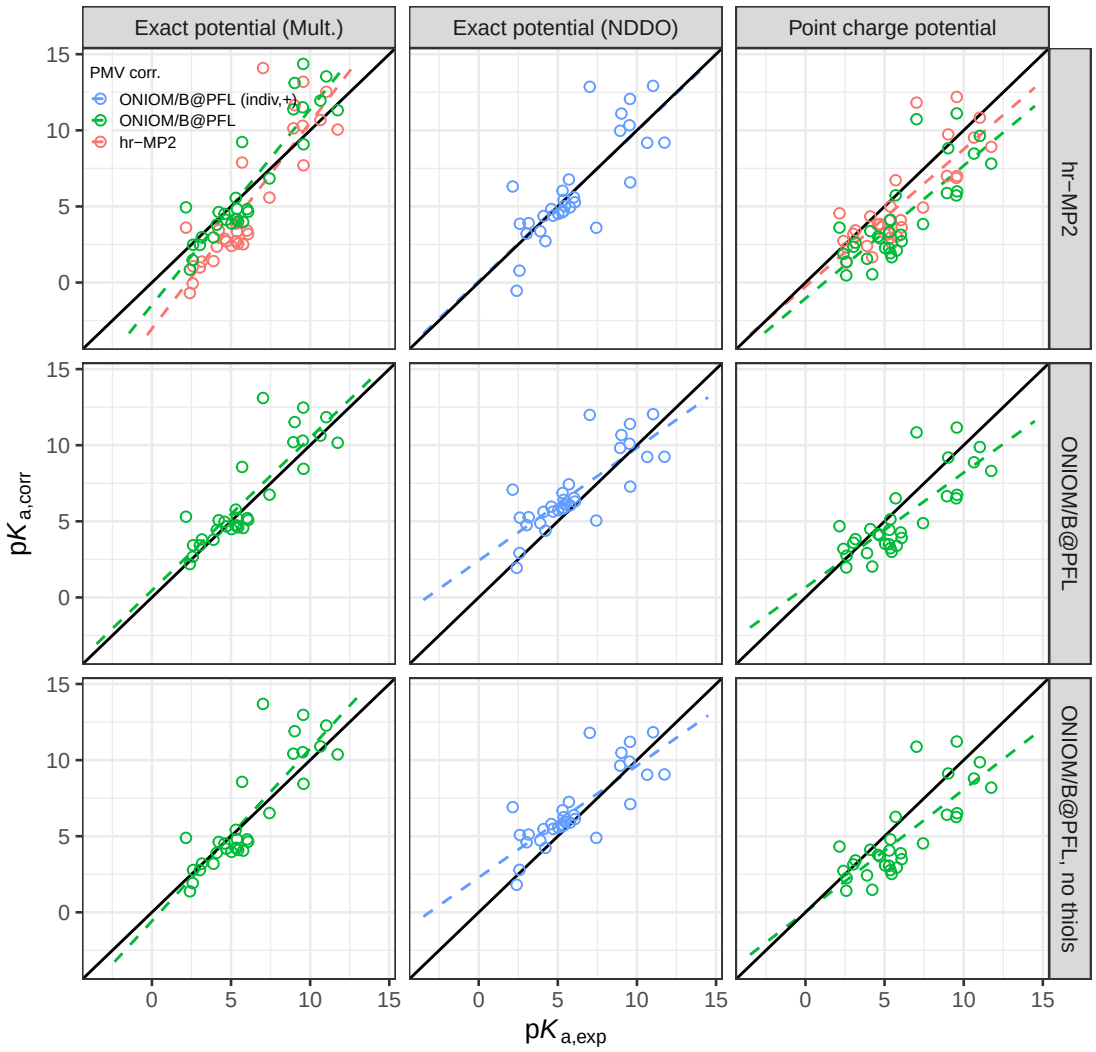

Figure 5.23.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A level of theory on PCM conformers reoptimised at the PM6-PCM level of theory. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv.,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

Table 5.21.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/A single point calculations on PM6-PCM-optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6 and 5.14 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. All PMV parameters used for the models in this table can be found in table 5.4, and pK_a correction parameters can be found in table 5.10 or table 5.12 for the thiol-free parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	2.51	2.21	-0.89	1.35	-3.04	0.74
	/B@PFL (All)	*hr*-MP2	/B|P|M|A‖*hr*|A	2.30	1.58	0.50	1.32	-1.43	0.75
	/B@PFL (All)	/B@PFL	/B|P|M|A‖A	1.65	1.09	0.48	1.00	0.46	0.75
	/B@PFL (All)	/B@PFL, n.t.	/B|P|M|A‖A|nt	1.83	1.28	0.22	1.13	-0.58	0.75
Exact (NDDO)	/B@PFL (Indiv.,+)	*hr*-MP2	/B|P|N|I+‖*hr*|A	1.98	1.46	0.02	0.99	0.10	0.64
	/B@PFL (Indiv.,+)	/B@PFL	/B|P|N|I+‖A	1.85	1.43	0.85	0.74	2.42	0.64
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|P|N|I+‖A|nt	1.77	1.32	0.68	0.73	2.29	0.64
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|P|A‖A	1.99	1.66	-0.85	0.90	-0.26	0.65
	/B@PFL (All)	*hr*-MP2	/B|P|P|A‖*hr*|A	2.56	2.25	-1.82	0.87	-1.04	0.64
	/B@PFL (All)	/B@PFL	/B|P|P|A‖A	1.87	1.57	-0.83	0.75	0.66	0.64
	/B@PFL (All)	/B@PFL, n.t.	/B|P|P|A‖A|nt	2.07	1.74	-1.16	0.80	0.03	0.64
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

predicted pK_a values, thus reducing the overall prediction quality.

5.3.7. Acidity constant prediction with ONIOM-EC-RISM/X

The last ONIOM-EC-RISM approximation to be evaluated is the /X scheme. As already described in section 4.6.4, this scheme allows for two PMV correction modes. In the first, /Xa, the ONIOM extrapolation is performed first and the resulting free energy is then corrected with a PMV correction scheme similar to those used for the other ONIOM-EC-RISM approximations. Therefore, the PMV and pK_a correction parameter combinations tested for the /A approximation are also applied to this scheme.

In the second scheme, /Xb, the PMV correction is performed first and then the ONIOM extrapolation is performed in a second step. Therefore, the set of PMV correction parameters is different from the other scheme. In principle, each free energy from the partial calculations can be corrected with its own set of parameters. However, since both the *lr* and *lm* corrections share the same level of theory, in this case PM6, they are corrected with the same set of parameters. These are either the PMV correction models obtained from the *lr*-PM6 level of theory using a single set of parameters for all charge states of the MNSOL data set, or alternatively the individual correction "*lr*-PM6 (indiv.,+)". In the latter case, the parameters obtained for the cationic subset are also

Table 5.22.: Definition of PMV corrections for the /Xb scheme. In all cases, a *hr*-EC-RISM correction from the SAMPL6 publication by Tielker et al.[24] is combined with a *lr*-EC-RISM correction parameterised in section 5.1. Thus, a combined PMV model ID is given in the last column. All *hr*parameterisations were obtained from B3LYP-PCM geometries and using a single set of parameters for all charge states. Further information on the individual model IDs can be found in tables 5.1 and 5.14. The PMV correction approach "Indiv.,+" is obtained by applying the cationic parameters to all ionic species and is marked as "I+" in the model IDs. In the main text and other tables in this chapter, the /Xb correction models are abbreviated as "*hr*-MP2 (All) & *hr*-MP2 (All)" and "*hr*-MP2 (All) & *hr*-MP2 (Indiv.,+)" respectively. See table 5.5 for the *hr*-PMV correction parameters.

hr PMV correction model		*lr* PMV correction model					PMV model ID
Potential	Model ID	Optimisation	Potential	Charge fit	Model ID	Tab.	
Exact[24]	*hr*|B|E|A	B3LYP-PCM	Exact (Mult.)	All	*lr*|B|M|A	(5.3)	*hr*&*lr*|B|M|A
				Indiv.	*lr*|B|M|I	(5.3)	*hr*&*lr*|B|M|A&I
			Exact (NDDO)	All	*lr*|B|N|A	(5.3)	*hr*&*lr*|B|N|A
				Indiv.	*lr*|B|N|I	(5.3)	*hr*&*lr*|B|N|A&I
		PM6-PCM	Exact (Mult.)	All	*lr*|P|M|A	(5.2)	*hr*&*lr*|P|M|A
				Indiv.	*lr*|P|M|I	(5.2)	*hr*&*lr*|P|M|A&I
			Exact (NDDO)	All	*lr*|P|N|A	(5.2)	*hr*&*lr*|P|N|A
				Indiv.	*lr*|P|N|I	(5.2)	*hr*&*lr*|P|N|A&I
Point charge[24]	*hr*|B|P|A	B3LYP-PCM	Point charge	All	*lr*|B|P|A	(5.3)	*hr*&*lr*|B|P|A
				Indiv.	*lr*|B|P|I	(5.3)	*hr*&*lr*|B|P|A&I
		PM6-PCM	Point charge	All	*lr*|P|P|A	(5.2)	*hr*&*lr*|P|P|A
				Indiv.	*lr*|P|P|I	(5.2)	*hr*&*lr*|P|P|A&I

Table 5.23.: Overview of the pK_a correction models for the /Xb correction scheme. The first column shows the PMV fit abbreviations that are used throughout this section. To reduce the total number of columns in this table, the details of the *hr*-PMV correction are not shown. However, in all cases, the *hr*-PMV correction using the exact potential is applied when an exact potential is applied for the *low* level PMV correction. The analogous case is true for the point charge potential. It should be noted that all details of the PMV correction can be retrieved using the provided PMV model IDs and table 5.22. For the given pK_a correction models, the corresponding corrections obtained from the training data set without thiols are marked with "|nt" as before, but are not shown explicitly. Their pK_a parameters can be found in tables 5.13 and 5.12. To limit the width of some columns, "Indiv." parameterisation schemes are abbreviated to "I", while "*hr*-MP2 (All)" is abbreviated to "*hr*-MP2". The associated pK_a correction parameters can be found in the tables shown next to the pK_a model IDs.

PMV corr.	*lr* PMV correction model				pK_a correction model			Mixed model ID
	lr-Optimisation	*lr*-Potential	Charge fit	Model ID	EC-RISM	Model ID	Tab.	
hr-MP2 & *lr*-PM6 (All)	B3LYP-PCM	Exact (Mult.)	All	*hr&lr*|B|M|A	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|M|A‖A	(5.11) (5.11)	*hr&lr*|B|M|A‖*hr*|A *hr&lr*|B|M|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	B3LYP-PCM	Exact (Mult.)	Indiv.	*hr&lr*|B|M|I	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|M|I‖A	(5.11) (5.11)	*hr&lr*|B|M|I‖*hr*|A *hr&lr*|B|M|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	B3LYP-PCM	Exact (Mult.)	Indiv.,+	*hr&lr*|B|M|I+	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|M|I+‖A	(5.11) (5.11)	*hr&lr*|B|M|I+‖*hr*|A *hr&lr*|B|M|I+‖/B|A
hr-MP2 & *lr*-PM6 (All)	B3LYP-PCM	Exact (NDDO)	All	*hr&lr*|B|N|A	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|N|A‖A	(5.11) (5.11)	*hr&lr*|B|N|A‖*hr*|A *hr&lr*|B|N|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	B3LYP-PCM	Exact (NDDO)	Indiv.	*hr&lr*|B|N|I	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|N|I‖A	(5.11) (5.11)	*hr&lr*|B|N|I‖*hr*|A *hr&lr*|B|N|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	B3LYP-PCM	Exact (NDDO)	Indiv.,+	*hr&lr*|B|N|I+	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|B|N|I+‖A	(5.11) (5.11)	*hr&lr*|B|N|I+‖*hr*|A *hr&lr*|B|N|I+‖/B|A
hr-MP2 & *lr*-PM6 (All)	B3LYP-PCM	Point charge	All	*hr&lr*|B|P|A	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|B|P|A‖A	(5.11) (5.11)	*hr&lr*|B|P|A‖*hr*|A *hr&lr*|B|P|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	B3LYP-PCM	Point charge	Indiv.	*hr&lr*|B|P|I	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|B|P|I‖A	(5.11) (5.11)	*hr&lr*|B|P|I‖*hr*|A *hr&lr*|B|P|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	B3LYP-PCM	Point charge	Indiv.,+	*hr&lr*|B|P|I+	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|B|P|I+‖A	(5.11) (5.11)	*hr&lr*|B|P|I+‖*hr*|A *hr&lr*|B|P|I+‖/B|A
hr-MP2 & *lr*-PM6 (All)	PM6-PCM	Exact (Mult.)	All	*hr&lr*|P|M|A	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|M|A‖A	(5.10) (5.10)	*hr&lr*|P|M|A‖*hr*|A *hr&lr*|P|M|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	PM6-PCM	Exact (Mult.)	Indiv.	*hr&lr*|P|M|I	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|M|I‖A	(5.10) (5.10)	*hr&lr*|P|M|I‖*hr*|A *hr&lr*|P|M|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	PM6-PCM	Exact (Mult.)	Indiv.,+	*hr&lr*|P|M|I+	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|M|I+‖A	(5.10) (5.10)	*hr&lr*|P|M|I+‖*hr*|A *hr&lr*|P|M|I+‖/B|A
hr-MP2 & *lr*-PM6 (All)	PM6-PCM	Exact (NDDO)	All	*hr&lr*|P|N|A	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|N|A‖A	(5.10) (5.10)	*hr&lr*|P|N|A‖*hr*|A *hr&lr*|P|N|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	PM6-PCM	Exact (NDDO)	Indiv.	*hr&lr*|P|N|I	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|N|I‖A	(5.10) (5.10)	*hr&lr*|P|N|I‖*hr*|A *hr&lr*|P|N|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	PM6-PCM	Exact (NDDO)	Indiv.,+	*hr&lr*|P|N|I+	*hr*-MP2 /B@PFL	*hr*|B|E|A‖A /B|P|N|I+‖A	(5.10) (5.10)	*hr&lr*|P|N|I+‖*hr*|A *hr&lr*|P|N|I+‖/B|A
hr-MP2 & *lr*-PM6 (All)	PM6-PCM	Point charge	All	*hr&lr*|P|P|A	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|P|P|A‖A	(5.10) (5.10)	*hr&lr*|P|P|A‖*hr*|A *hr&lr*|P|P|A‖/B|A
hr-MP2 & *lr*-PM6 (I)	PM6-PCM	Point charge	Indiv.	*hr&lr*|P|P|I	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|P|P|I‖A	(5.10) (5.10)	*hr&lr*|P|P|I‖*hr*|A *hr&lr*|P|P|I‖/B|A
hr-MP2 & *lr*-PM6 (I+)	PM6-PCM	Point charge	Indiv.,+	*hr&lr*|P|P|I+	*hr*-MP2 /B@PFL	*hr*|B|P|A‖A /B|P|P|I+‖A	(5.10) (5.10)	*hr&lr*|P|P|I+‖*hr*|A *hr&lr*|P|P|I+‖/B|A

applied to the anionic subset. In the following tables these models are marked with an ampersand sign, e.g. "*hr*-MP2 (All) & *lr*-PM6 (All)" indicates that the PMV correction is performed first and that the *hr*-MP2 parameters are used to correct the *hr* free energy, while the corresponding *lr*-PM6 parameters are applied to both *lr* and *lm* results.

Tables 5.22 and 5.23 provide an overview of the PMV and pK_a correction models for the /Xb correction scheme, analogous to tables 5.1, 5.6, 5.7 and 5.14. As both a *lr* and a *hr* correction are combined to give the /Xb PMV correction, the result is a mixed model, marked "*hr*&*lr*" in the model ID. As a consequence, the pK_a correction of /Xb is also a mixed model and is labelled using an analogous scheme to that used in table 5.14.

B3LYP-PCM geometries

Figure 5.24 shows the pK_a prediction results from the first scheme, while figure 5.25 shows the corresponding results for the second scheme. The statistical analysis of all /X models evaluated on the B3LYP-PCM geometries can be found in table 5.24. The corresponding predicted pK_a values are shown in tables 75 to 83 in the Appendix.

Surprisingly, most of the multipole-based models evaluated here show an RMSE that is smaller than that of the *hr*-reference model. The best prediction quality can be obtained from the "ONIOM first" scheme (/Xa) and the parameter combinations that also worked well for the /A scheme. These are the "/B@PFL (All)" PMV correction in combination with the /B@PFL pK_a correction parameterised with (/B|B|M|A‖A) or without (/B|B|M|A‖A|nt) the thiol subset, giving an excellent RMSE of 0.73 and 0.78 respectively, and thus predictions equal to or slightly better than the best performing /A model. The remaining two models applying the ONIOM extrapolation first, both based on the *hr*-MP2 pK_a correction parameters, give an RMSE of 1.15 and 1.37 or greater and are therefore worse than the *hr*-reference model, although the difference is small. This is consistent with the results obtained for the /A model. It is noteworthy that for all these models a high R^2 of 0.93 or greater is obtained.

For the second set of multipol-based models, which perform the ONIOM extrapolation in a second step (/Xb) and therefore after the PMV correction of the individual sub-calculations, the best model gives an RMSE of 0.99. This model was obtained by applying the "*lr*-PM6 (indiv.,+)" PMV correction to the *low* level calculations and the *hr*-MP2 pK_a correction parameters (*hr*&*lr*|B|M|I+‖*hr*|A). Although this is the best model of this kind and gives better predictions than the more expensive *hr*-reference, it gives a higher RMSE than the best /B model. The other "ONIOM second" models also give RMSE values higher than this more approximate level of theory.

As for the ONIOM-EC-RISM/A and /B schemes, the NDDO models generally give worse results than the corresponding multipole based models. In contrast to these multipole models, the "ONIOM second" (/Xb) approach generally gives better results, when compared to the "ONIOM first" (/Xa). The best overall NDDO-results are obtained with the "*hr*-MP2 (All) & *lr*-PM6 (All)" PMV correction and the /B@PFL pK_a correc-

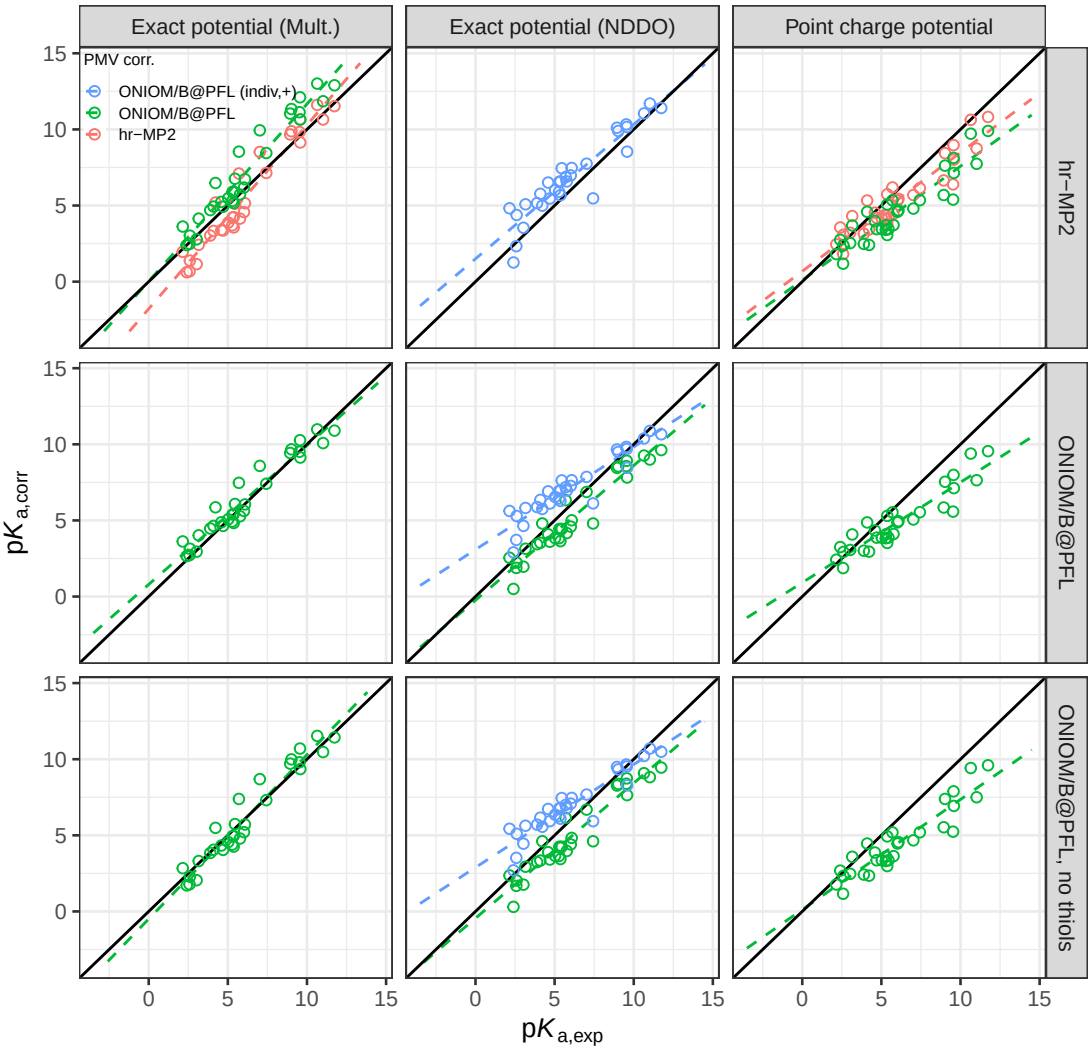

Figure 5.24.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory.[24] To avoid overplotting, only PMV corrections are shown where the ONIOM extrapolation is performed first (/Xa). The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv.,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

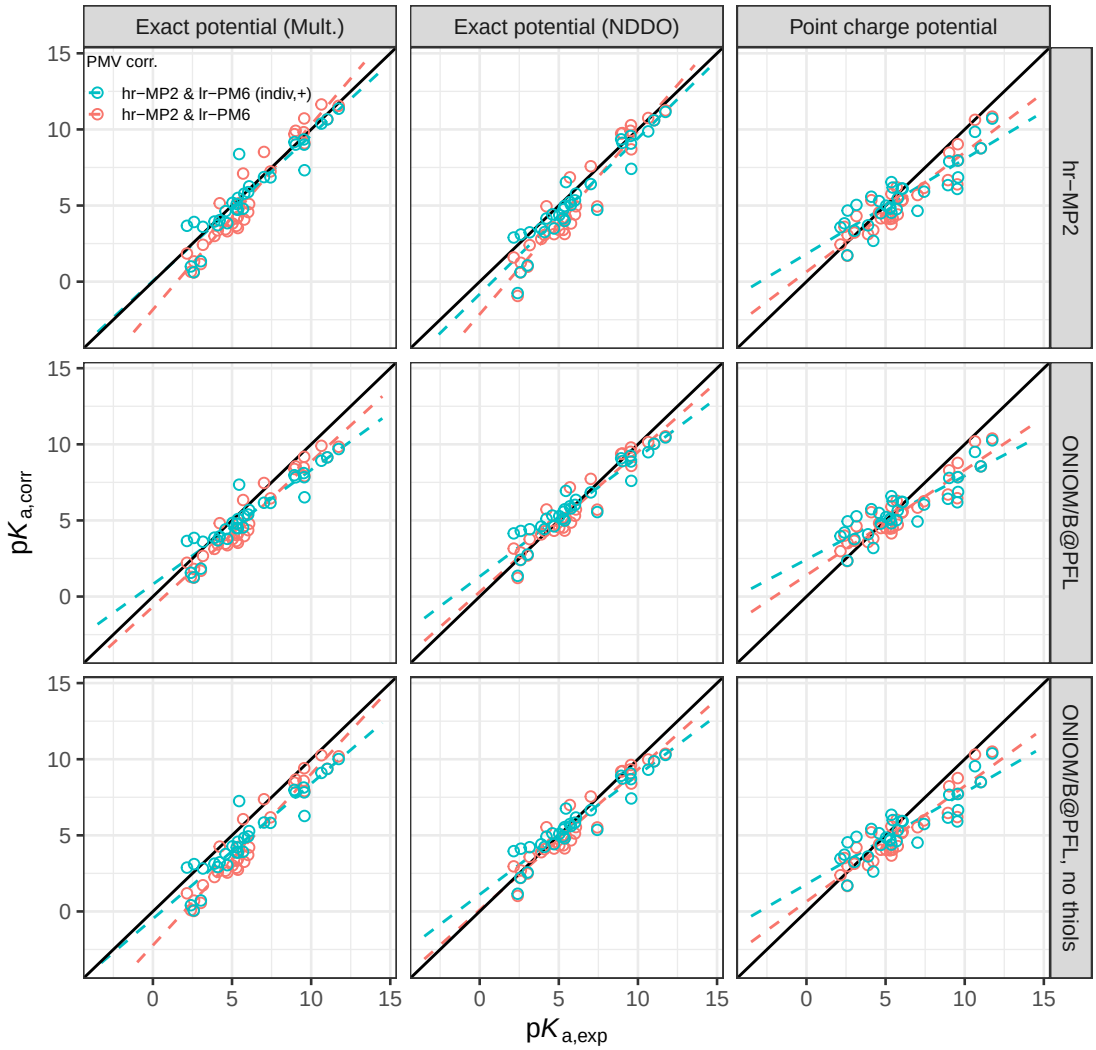

Figure 5.25.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers optimised at the B3LYP/6-311+G** level of theory.[24] To avoid overplotting, only PMV corrections are shown where the ONIOM extrapolation is performed second (/Xb). The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "*hr*-MP2 (All) & *lr*-PM6 (indiv.,+)" (blue) and "*hr*-MP2 (All) & *lr*-PM6 (All)" (red) parameters.

Table 5.24.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X single point calculations on B3LYP-PCM-optimised geometries.[24] The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6, 5.14, 5.22 and 5.23 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. The "*hr*-MP2 (All) & *lr*-PM6 (Indiv.,+)" PMV correction is also abbreviated to "*hr*-MP2 & *lr*-PM6 (I+)" for brevity. Please refer to tables 5.14, 5.22 and 5.23 for an overview of where to find the associated PMV and pK_a correction parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.15	1.04	-0.53	1.21	-1.79	0.93
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	1.37	1.06	1.02	1.16	0.03	0.94
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	0.73	0.54	0.27	0.91	0.80	0.94
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	0.78	0.65	-0.02	1.08	-0.49	0.94
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|M|A‖*hr*|A	1.17	1.05	-0.54	1.21	-1.84	0.93
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|M|A‖/B|A	1.18	1.06	-0.95	0.95	-0.67	0.93
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|M|A‖/B|A|nt	1.70	1.51	-1.46	1.13	-2.22	0.93
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|M|I+‖*hr*|A	0.99	0.67	-0.22	0.96	0.04	0.88
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|M|I+‖/B|A	1.22	1.02	-0.69	0.75	0.81	0.88
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|M|I+‖/B|A|nt	1.49	1.36	-1.16	0.89	-0.48	0.88
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|B|N|A‖A	1.21	1.04	-0.94	0.88	-0.24	0.92
	/B@PFL (All)	/B@PFL, n.t.	/B|B|N|A‖A|nt	1.36	1.19	-1.13	0.89	-0.45	0.92
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.24	1.10	0.79	0.88	1.50	0.87
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.58	1.37	1.12	0.67	3.08	0.87
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.45	1.25	0.93	0.68	2.89	0.87
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|N|A‖*hr*|A	1.42	1.24	-0.91	1.20	-2.14	0.92
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|N|A‖/B|A	0.78	0.64	-0.18	0.92	0.30	0.92
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|N|A‖/B|A|nt	0.84	0.69	-0.37	0.92	0.10	0.92
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|N|I+‖*hr*|A	1.15	0.86	-0.67	1.02	-0.82	0.90
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|N|I+‖/B|A	0.92	0.69	-0.00	0.78	1.32	0.90
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|N|I+‖/B|A|nt	0.94	0.71	-0.19	0.78	1.11	0.90
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.20	0.99	-0.64	0.78	0.68	0.87
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	1.78	1.51	-1.42	0.75	0.10	0.85
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.64	1.35	-1.14	0.66	0.92	0.85
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	1.90	1.62	-1.55	0.72	0.12	0.85
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|P|A‖*hr*|A	1.21	1.00	-0.65	0.78	0.65	0.87
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|P|A‖/B|A	1.20	1.00	-0.47	0.69	1.40	0.87
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|P|A‖/B|A|nt	1.31	1.10	-0.81	0.76	0.65	0.87
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|P|I+‖*hr*|A	1.46	1.18	-0.45	0.62	1.83	0.76
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|P|I+‖/B|A	1.50	1.20	-0.29	0.55	2.44	0.76
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|P|I+‖/B|A	1.54	1.25	-0.61	0.60	1.79	0.76
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

tion (*hr*&*lr*|B|N|A‖/B|A), giving an RMSE of 0.78, which is a prediction quality close to the best multipole model.

The application of the point charge ESP in the ONIOM-EC-RISM/X scheme generally produces models that are better than the corresponding /A models. Here the best RMSE values of 1.20 for all point charge models are observed for two models. The first is obtained by the application of both *hr* corrections to the "ONIOM first" approach (*hr*|B|E|A||A). The second is obtained through the application of the "*hr*-MP2 & lr-PM6 (All)" PMV correction and the /B@PFL pK_a correction. In addition, a slightly worse RMSE of 1.21 is obtained by using the *hr*-MP2 pK_a correction instead of the /B@PFL pK_a correction. Again changing the set of pK_a correction parameters to those obtained from the training data set without thiols, gives an RSME of 1.31. Other than that, no significant models can be observed as the prediction quality is generally worse than the *hr* reference and the more cost-effective /B scheme.

ONIOM-PCM/X geometries

As before, the effect of the approximate ONIOM-PCM/X optimisation level on the prediction quality of the ONIOM-EC-RISM/X schemes is also investigated. It should be emphasised that the level of optimisation and the level of theory of the single point calculation that are now used are conceptually very similar. Both ONIOM solvation models have in common that the free energy of the *hr* calculation is approximated on the basis of three independent calculations, as already described in detail in sections 3.3 and 4.6.

The results of the "ONIOM first" approach are shown in figure 5.26, while the results of the "ONIOM second" approach are shown in 5.27. The statistical quantities for both schemes are given in table 5.25. The corresponding predicted pK_a values are given in tables 84 to 92 in the appendix.

The similarity between the results obtained from the /A and /X schemes is also evident from the effect of the reduced level of optimisation. For the multipole based /A scheme it was observed that the application of the ONIOM-PCM/X reoptimised structures leads to a systematic shift towards smaller predicted pK_a values. The same effect can be observed for the /X scheme, as indicated by the decrease in the MSE for all these models. Although the effect is small, it leads to an increase or decrease in the prediction quality, depending on whether the B3LYP-PCM structures lead to an overprediction or underprediction of the experimental pK_a values. In other words, the error introduced by the application of the reoptimised structures may directly compensate for the overprediction observed for the less approximate level of optimisation.

This is the case for the already well performing model resulting from the application of the "/B@PFL (All)" PMV correction and the /B@PFL pK_a correction (/B|B|M|A‖A), i.e. the combination of parameters that also gave excellent results for the /A scheme and good results for the /B scheme. Here, the usage of the ONIOM-PCM/X conformers leads

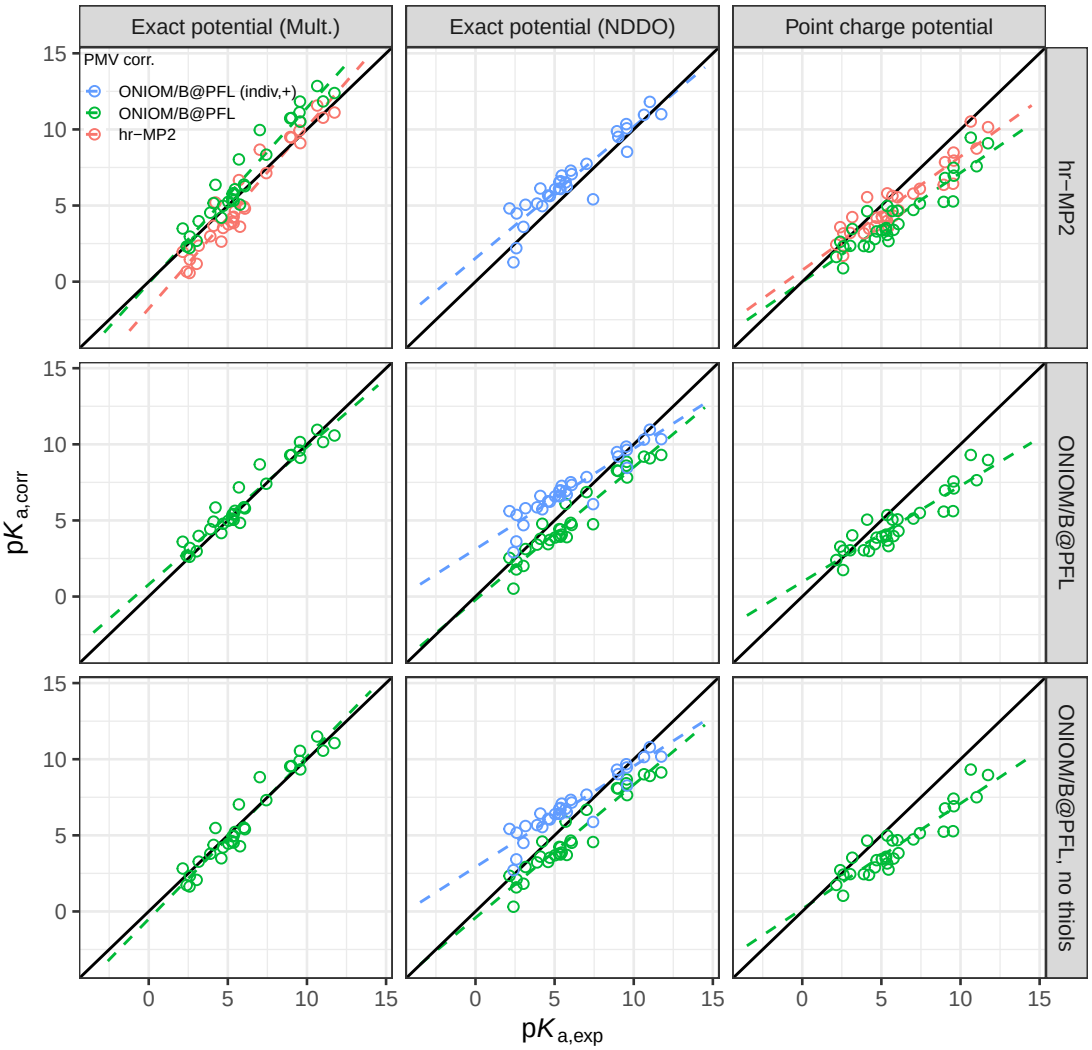

Figure 5.26.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers reoptimised at the ONIOM-PCM/X level of theory. To avoid overplotting, only PMV corrections are shown where the ONIOM extrapolation is performed first (/Xa). The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "/B@PFL (indiv,+)" (blue), "/B@PFL" (green) and "*hr*-MP2" (red) parameters.

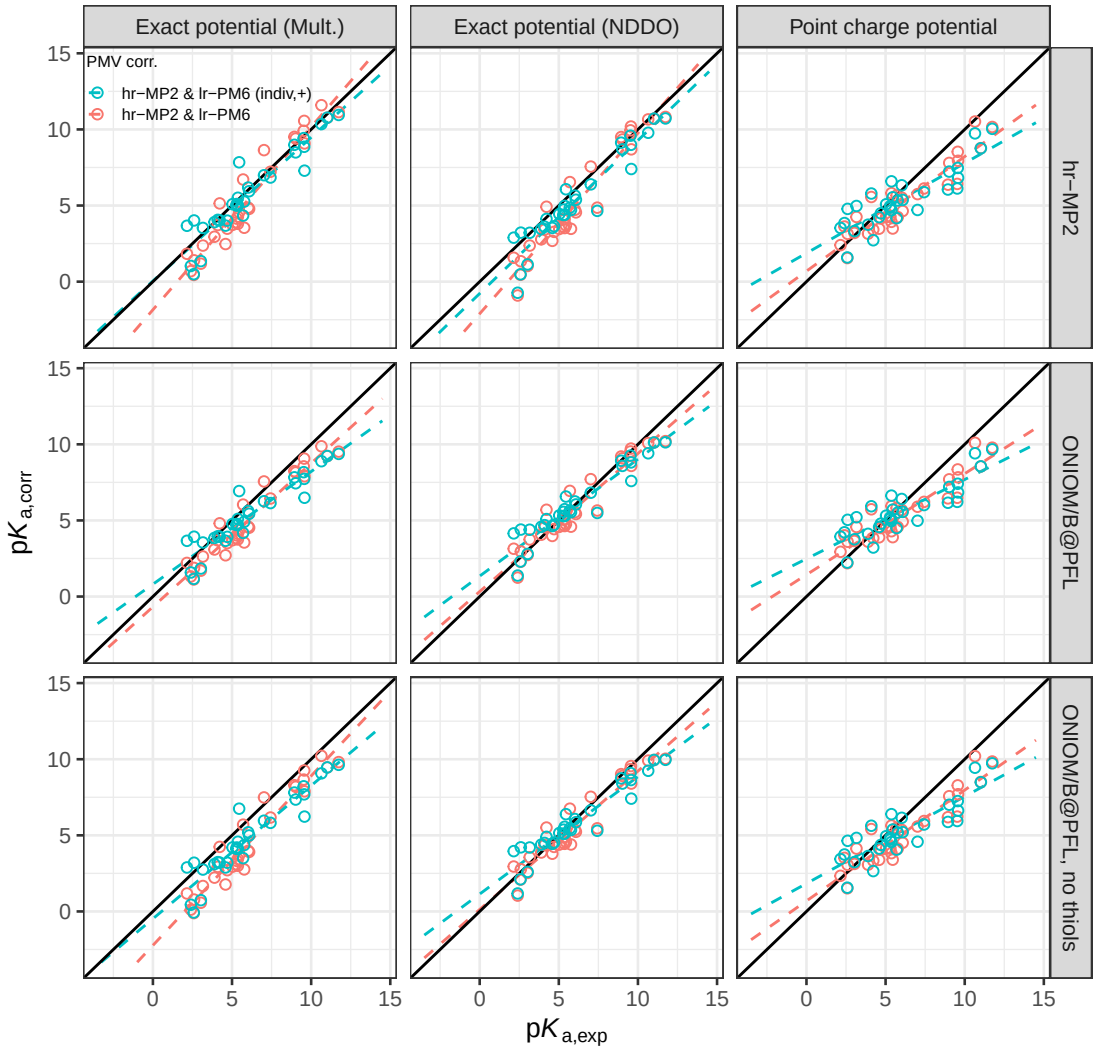

Figure 5.27.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers reoptimised at the ONIOM-PCM/X level of theory. To avoid overplotting, only PMV corrections are shown where the ONIOM extrapolation is performed second (/Xb). The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text. The colours indicate PMV corrections based on the "*hr*-MP2 (All) & *lr*-PM6 (indiv.,+)" (blue) and "*hr*-MP2 (All) & *lr*-PM6 (All)" (red) parameters.

to a slight decrease in the MSE from 0.27 to 0.21. Consequently, the RMSE decreases from 0.73 to 0.72, giving the best prediction for the SAMPL6 dataset from any of the ONIOM-EC-RISM models. The same effect can be observed for the model using the corresponding thiol-free set of pK_a correction parameters (/B|B|M|A‖A|nt). However, it should be stressed that the increase in prediction quality is minimal in both cases. At the same time, there is no multipole based model where a drastic decrease in prediction quality is observed.

The same reduction in MSE can also be observed for the NDDO and point charge based models, except for the NDDO model using the "ONIOM second" approach and the "*hr*-MP2 (All) & *lr*-PM6 (indiv,+)" PMV correction and the /B@PFL pK_a correction. Here the MSE increases slightly.

In general, the NDDO models that performed well on the B3LYP-PCM geometries also give good results on the reoptimised structures, but still underperform compared to the multipole based ESP approximations. The same is true for the point charge models, where again no significant models are obtained.

The results so far indicate that the ONIOM-EC-RISM/X model gives predictions very similar to the /A results, albeit with a slightly higher prediction quality. For the /A model it could be seen that the PM6-PCM conformers do not give usable results. This is also true for the /X model. For this reason, these models are not discussed and the corresponding data are only presented in tables 15 and 93 to 101 as well as figures 5 and 6 in the appendix.

Table 5.25.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X single point calculations on ONIOM-PCM/X optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6, 5.14, 5.22 and 5.23 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. The "*hr*-MP2 (All) & *lr*-PM6 (Indiv.,+)" PMV correction is also abbreviated to "*hr*-MP2 & *lr*-PM6 (I+)" for brevity. Please refer to tables 5.14, 5.22 and 5.23 for an overview of where to find the associated PMV and pK_a correction parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.18	1.05	-0.60	1.19	-1.77	0.93
	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	1.22	0.94	0.81	1.15	-0.10	0.94
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	0.72	0.52	0.21	0.90	0.79	0.94
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	0.78	0.66	-0.10	1.07	-0.50	0.94
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|M|A‖*hr*|A	1.21	1.07	-0.63	1.20	-1.85	0.93
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|M|A‖/B|A	1.24	1.12	-1.02	0.94	-0.67	0.93
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|M|A‖/B|A|nt	1.77	1.58	-1.55	1.11	-2.23	0.93
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|M|I+‖*hr*|A	0.99	0.69	-0.30	0.94	0.04	0.88
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|M|I+‖/B|A	1.26	1.06	-0.76	0.74	0.81	0.88
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|M|I+‖/B|A|nt	1.55	1.41	-1.24	0.87	-0.48	0.88
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|B|N|A‖A	1.25	1.08	-0.99	0.87	-0.21	0.92
	/B@PFL (All)	/B@PFL, n.t.	/B|B|N|A‖A|nt	1.40	1.23	-1.18	0.87	-0.42	0.92
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|B|N|I+‖*hr*|A	1.19	1.06	0.72	0.87	1.54	0.88
	/B@PFL (Indiv.,+)	/B@PFL	/B|B|N|I+‖A	1.55	1.34	1.06	0.66	3.11	0.87
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|B|N|I+‖A|nt	1.43	1.22	0.88	0.66	2.92	0.87
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|N|A‖*hr*|A	1.44	1.24	-0.98	1.19	-2.11	0.92
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|N|A‖/B|A	0.79	0.65	-0.24	0.91	0.33	0.92
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|N|A‖/B|A|nt	0.86	0.72	-0.42	0.91	0.13	0.92
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|N|I+‖*hr*|A	1.17	0.89	-0.74	1.01	-0.78	0.90
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|N|I+‖/B|A	0.93	0.67	-0.06	0.77	1.34	0.90
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|N|I+‖/B|A|nt	0.95	0.72	-0.24	0.77	1.14	0.90
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.31	1.11	-0.78	0.75	0.75	0.86
	/B@PFL (All)	*hr*-MP2	/B|B|P|A‖*hr*|A	2.05	1.79	-1.72	0.72	-0.02	0.85
	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	1.76	1.49	-1.27	0.63	0.96	0.85
	/B@PFL (All)	/B@PFL, n.t.	/B|B|M|A‖A|nt	2.04	1.77	-1.69	0.69	0.17	0.85
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|B|P|A‖*hr*|A	1.33	1.14	-0.80	0.75	0.69	0.86
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|B|P|A‖/B|A	1.31	1.10	-0.60	0.66	1.43	0.86
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|B|P|A‖/B|A|nt	1.44	1.24	-0.95	0.73	0.69	0.86
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|B|P|I+‖*hr*|A	1.56	1.27	-0.60	0.59	1.87	0.74
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|B|P|I+‖/B|A	1.58	1.27	-0.42	0.52	2.48	0.74
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|B|P|I+‖/B|A	1.65	1.35	-0.75	0.57	1.84	0.74
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

5.3.8. Overview of the best performing ONIOM-EC-RISM models

A large number of ONIOM-EC-RISM pK_a prediction models have been presented so far. In order to provide a brief overview and conclusion for the method validation on the SAMPL6 dataset, a summary of the best performing models is given below. For this purpose, a selection of models is provided in table 5.26. The multipole approximation to the exact potential was used to derive the models in all cases and the following discussion is limited to this ESP-approximation.

Table 5.26.: Selection of the best performing ONIOM-EC-RISM models compared to the best *hr*-reference model,[24] which is given in the last row. All ONIOM-EC-RISM models were obtained from the multipole ESP approximation. The first set of models highlights the increase in prediction quality when decreasing the level of ONIOM-EC-RISM approximation from /B to /A and /X, while keeping the set of correction parameters constant. The second set of models shows the overall best models for the respective ONIOM-EC-RISM approximation, in case they were not included in the first set. Note that the two best-performing /X models shown here were obtained using the "ONIOM first" (/Xa) PMV correction scheme. See tables 5.14, 5.23 and the corresponding sections for an overview of the Model IDs. The columns next to the model ID show the tables containing the corresponding PMV and pK_a correction parameters. For context, the tables containing the evaluation on the SAMPL6 dataset are also provided.

EC-RISM	Optimisation	PMV fit	pK_a fit	Model ID	Tables			RMSE	MAE	MSE	m'	b'	R^2
					PMV	pK_a	SAMPL6						
/B@PFL	B3LYP-PCM	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	(5.5)	(5.11)	(5.18)	1.17	0.90	-0.25	0.69	1.59	0.85
/B	B3LYP-PCM	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	(5.5)	(5.11)	(5.15)	1.13	0.89	-0.61	0.73	0.99	0.91
/A	B3LYP-PCM	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	(5.5)	(5.11)	(5.19)	0.78	0.59	0.06	0.92	0.56	0.92
/X	B3LYP-PCM	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	(5.5)	(5.11)	(5.24)	0.73	0.54	0.27	0.91	0.80	0.94
/B@PFL	B3LYP-PCM	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	(5.5)	(5.11)	(5.18)	1.12	0.75	0.35	0.89	1.04	0.85
/B	B3LYP-PCM	/B@PFL (All)	*hr*-MP2	/B|B|M|A‖*hr*|A	(5.5)	(5.11)	(5.15)	0.85	0.71	-0.16	0.93	0.29	0.90
/X	ONIOM-PCM/X	/B@PFL (All)	/B@PFL	/B|B|M|A‖A	(5.5)	(5.11)	(5.25)	0.72	0.52	0.21	0.90	0.79	0.94
hr-ref.[24]	B3LYP-PCM	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	(5.5)	(5.11)	(5.15)	1.13	0.97	-0.36	1.17	-1.38	0.91

The best performing models can be divided into two groups: Firstly, all models resulting from the application of the "/B@PFL (All)" PMV correction and the /B@PFL pK_a correction (/B|B|M|A‖A), as they give consistently good results. Second, the models that gave the best overall predictions for the respective ONIOM-EC-RISM schemes, if they were not already included in the first set.

Note that this means that in this second set, different pK_a correction parameters are used for the ONIOM-EC-RISM/B and /B@PFL schemes compared to the first set, while the best performing /X model shown uses a different set of geometries.

From this first set, the ONIOM-EC-RISM/B@PFL model gives an RMSE of 1.17. Its prediction quality is therefore more or less equal to the more expensive *hr*-reference model given in the last row of table 5.26, which resulted in a value of 1.13 pK_a units. By evaluating this model at the size extrapolation limit, i.e. the partition-free limit (PFL),

the pK_a predictions are free of any ONIOM partitioning effect.

The partitioning of the SAMPL6 molecules for the /B scheme gives a slightly improved RMSE of 1.13, while further reducing the cost of the EC-RISM calculation by excluding parts of the molecules from the expensive *high* level of theory. From the analysis of the partitioning effect, it can be concluded that partitioning leads to a cancellation of the errors observed for the /B model or, more generally, for the EC-RISM solvation model, thus increasing the overall prediction quality of the model. This effect is more pronounced for the /B models using the *hr*-MP2 pK_a correction in table 5.26.

Although the prediction of the /B model is already equal to that of the *hr* reference model on this test data set, reducing the level of approximation of the ONIOM-EC-RISM method leads to a significant improvement in the predictions. Applying the /A model with the same set of correction parameters as before improves the RMSE to a value of 0.78 with an R^2 of 0.92. This is the overall best /A model that could be obtained.

A further improvement can be achieved by switching to the /X scheme, more precisely the "ONIOM first" scheme (/Xa). This results in a further reduction of the RMSE to a value of 0.73, while R^2 increases to 0.94. This is the best overall result for the first set of models obtained from the B3LYP-PCM geometries and the /B@PFL corrections. This excellent result can only be further improved by applying the reoptimised ONIOM-PCM/X geometries, which gives an RSME of 0.72, although this improvement is minimal.

This slight improvement is most likely due to a fortunate cancellation of errors resulting from a systematic downshift of the predicted pK_a values that can be observed for all /X models when the optimisation level is reduced. A similar effect is observed for the /B model, which uses the *hr*-pK_a correction instead of the consistently parameterised /B@PFL correction. As explained in detail above, the parameter combination used here, in particular the additive effect of c_q and m, coincidentally leads to error cancellation, resulting in a very good RMSE of 0.85.

Furthermore, it is noteworthy that all of the models that performed best, as shown in this section, use the classical correction scheme employed in the *hr*-reference publication by Tielker et al.,[24] rather than the individual corrections that performed well on the training dataset. It is therefore likely that their improved performance on the training dataset was due to overfitting. While this work has found models that perform excellently on the test dataset and the employed statistical analysis is in line with the reference publication, subsequent studies should revisit the parameterisation of the two correction models using more sophisticated statistical methods, such as cross-validation, to investigate the potential overfitting of the individual corrections.

From this summary it can be seen that reducing the level of approximation of the ONIOM-EC-RISM schemes, while keeping the applied correction parameters constant, leads to a significant reduction in the RMSE and thus to a drastic improvement in the overall prediction quality. Additionally, in the worst case, all the models in table 5.26 give predictions that are essentially equal to, and in most cases significantly better than,

the *hr*-reference model. In other words, even if the fortunate error cancellation observed for calculations on partitioned structures is removed by applying /B@PFL, it is still possible to obtain models that give a prediction quality comparable to that of the *hr*-EC-RISM model, while at the same time it can be expected that they will lead to a drastic reduction in costs, the specific amount of which will be measured in the next section.

5.3.9. Measuring the speed-up of the ONIOM-EC-RISM approximations

Following the validation that the ONIOM-EC-RISM schemes are capable of reproducing and outperforming the *hr*-EC-RISM scheme for the SAMPL6 data set, it will be demonstrated that the ONIOM schemes also offer a significant computational advantage in terms of costs, which was the primary motivation for their development.

Figure 5.28 shows the runtimes calculated for EC-RISM single point calculations on ONIOM-PCM/X reoptimised conformers for the four partitions of SM24 with the three ONIOM-EC-RISM schemes /A, /B and /X as well as the corresponding calculations at the *hr* and *lr* levels of theory as a reference.

As the ONIOM-PCM/X reoptimisation depends on the applied ONIOM partitioning, four different geometries were obtained, resulting in slightly different total run times for the *hr* and *lr* schemes, as shown in 5.28. All single point calculations were performed on identical machines equipped with Intel Xeon E5 2640v4 CPUs with 20 cores and 60 GB of memory. All calculations were performed using the exact ESP and, in the case of ONIOM-EC-RISM, the multipole approximation.

From figure 5.28 it is clear that two extreme cases can be identified based on the total runtimes. On the one hand, the *hr*-EC-RISM scheme shows runtimes that are in all cases greater than all other EC-RISM schemes. Most of this run time can be attributed to the electronic structure calculation at the *high* level of theory on the *real* system, since this calculation has to be repeated in each EC-RISM iteration. For the same reasons, the second major part of the total run time is caused by the ESP generation. In contrast, the calculation of the solvent structure by the 3D-RISM solver takes a relatively small part of the total run time.

On the other hand, the *lr*-EC-RISM scheme with the PM6 Hamiltonian drastically reduces these costs by minimising the cost of electronic structure calculations. Although this model is therefore extremely cost efficient, it has been shown previously that no reasonable free energy of solvation or pK_a predictions can be obtained from this level of theory. It is therefore only used here as a reference for the lower bound of the cost that can be achieved by reducing the level of theory.

In the ONIOM-EC-RISM/A scheme, the costly *hr* electronic structure calculation is approximated by an ONIOM extrapolation. As a consequence, three electronic structure calculations must be performed in each EC-RISM iteration, as well as the corresponding routines for ESP generation. This scheme will therefore only be able to reduce the

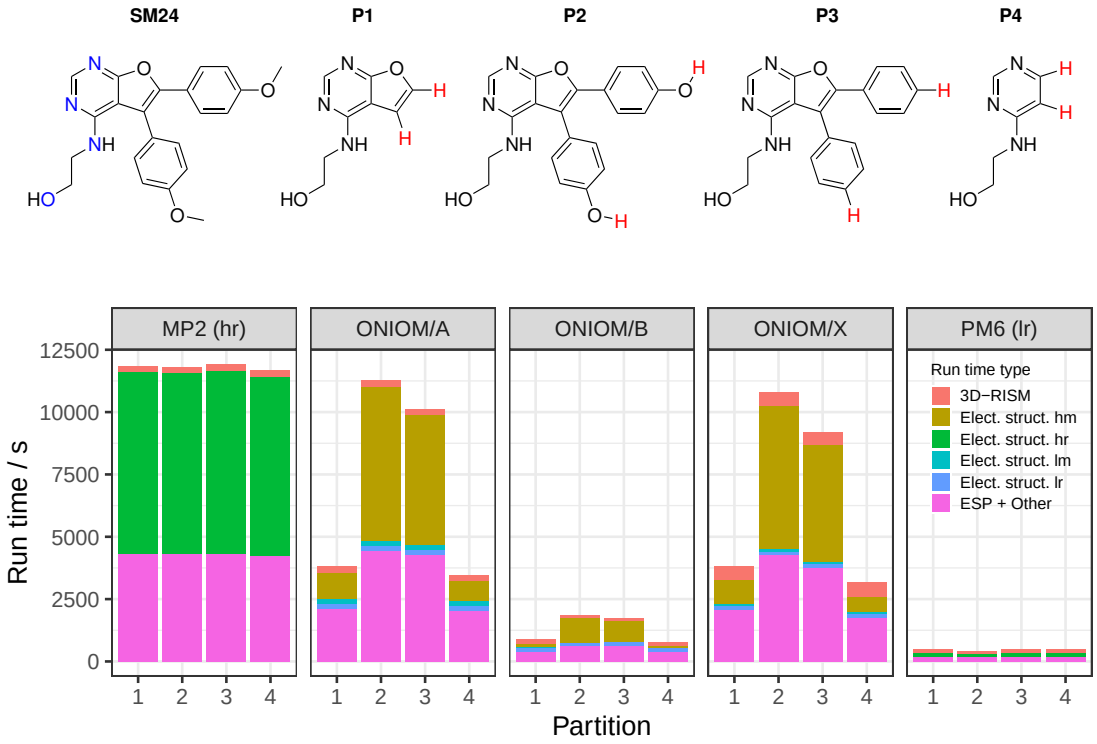

Figure 5.28.: Runtime measurements of different EC-RISM schemes and theory levels for the reoptimised ONIOM-PCM/X conformers and four partitions of SM24. Runtimes are broken down into individual contributions from solvent structure calculations with 3D-RISM, electronic structure calculations and ESP generation, as well as additional overhead caused by the script-based implementation of EC-RISM. The upper portion of the figure depicts the chemical structure and employed *model* systems for SM24.

cost compared to the *hr*-EC-RISM scheme if the sum of the costs associated with these sub-calculations is less than the time required to perform the *hr* calculation including the *hr*-ESP generation. As figure 5.28 shows, this strongly depends on the partitioning applied. For P2 and P3, which exclude only small parts of the molecule from the *model* system, the cost reduction is minimal and the cost of ESP generation and additional overhead actually increases compared to the *hr* scheme. For the other two partitions, P1 and P4, a more significant cost reduction is observed, which can be attributed to the smaller *model* systems generated by these partitions and thus the reduced costs for both the *hm* calculation and the ESP generation.

The same argument can be made for the ONIOM-EC-RISM/X scheme. Here, too, the expensive *high* level calculations must be repeated in each iteration resulting in similar

dependencies on the partitioning. However, it can be observed that the *low* level sub-calculations *lr* and *lm* tend to converge faster than the *hm* calculation. Since in the /X scheme all three sub-calculations are independent, this leads to a reduction in the total number of EC-RISM iterations and therefore to a reduction in the run time for the *lr* and *lm* electronic structure calculations. In addition to this effect, the independence of the sub-calculations leads to an additional potential speed-up, as all three sub-calculations can be run in parallel if sufficient computational resources are available. It should be noted, however, that this independence also implies the need to perform 3D-RISM calculations in each of the EC-RISM sub-calculations, which increases the contribution of the solvent structure calculation to the total run time.

In the ONIOM-EC-RISM/B scheme the solvent structure calculation is completely decoupled from the two *model* system calculations. As a consequence, the *lr*-ESP and the solvent environment can be converged first, before in a final iteration the *hm* and *lm* electronic energies have to be evaluated only once in the already converged solvent point charge field. This leads to a drastic reduction in the overall run time as shown in figure 5.28. The total cost is therefore identical to the cost of obtaining all the *model* system quantities and the cost of a *lr*-EC-RISM calculation. This also means that the strong dependence on *model* system size observed for the other ONIOM-EC-RISM schemes is removed. In light of the observed high prediction quality for the SAMPL6 data, it can be concluded that the /B scheme represents an effective means of achieving a balance between speed and accuracy.

Therefore, the choice of ONIOM-EC-RISM scheme for a specific research scenario depends on the required accuracy and size of the *model* system, as well as the number of structures to be investigated, i.e. the number of single-point calculations to be performed.

The /B model is reasonable for cases where a large number of structures need to be calculated, since its overall cost is the lowest of all the ONIOM-EC-RISM schemes. The scheme is also applicable in scenarios where large *model* systems are required, since only one expensive, *high* level electronic structure calculation is performed, thus requiring less computational effort than the other schemes. According to the results obtained in this chapter, B3LYP-PCM geometries and the /B@PFL PMV correction should be used in conjunction with the /B@PFL or *hr*-MP2 pK_a correction. Both produced results equal to or better than the *hr*-reference model (see table 5.26), although subsequent studies should perform further calculations on additional test data sets to identify a consistently well-performing parameter set.

If a higher level of precision is required than that offered by the /B scheme, either the /A or the /X scheme could be used, since both schemes demonstrate comparable prediction qualities on the SAMPL6 dataset (see table 5.26). However, the /X scheme enables subcalculations to be performed in parallel, and each subcalculation can converge freely, possibly requiring fewer EC-RISM iterations than the /A scheme. Therefore, the /X scheme is preferable to the /A scheme. Note that both schemes have a stronger scaling relationship with the *model* system size than the /B scheme. For the given method

combination, the "ONIOM first" (/Xa) PMV correction scheme using the /B@PFL parameters should be applied in combination with the /B@PFL pK_a parameters. Here, either B3LYP-PCM or ONIOM-PCM/X geometries can be used, depending on availability, as they produce similar prediction qualities.

6. Chemical shift and acidity constant prediction for a GEAEG pentapeptide

The validation on the SAMPL6 data set demonstrated that it is possible to use the ONIOM-EC-RISM method for the accurate and cost-efficient prediction of pK_a values for small molecules. While the observed accuracy was in most cases equal or better than *hr*-EC-RISM, the investigated system sizes still allowed for calculations with *hr*. As the initial motivation for the development of ONIOM-EC-RISM was to reduce the computational cost to allow the investigation of larger biomolecular systems that can no longer be treated with this *hr*-reference method, the size of the investigated system will be increased to explore the capabilities of the multiscale solvation model on larger systems. In addition, this provides a way to validate the model for medium to larger sized systems.

This additional validation requires not only the availability of experimental reference values, but also a comparison with already established computational methods, as the increased system sizes now prevent the use of the *hr*-EC-RISM method and therefore a direct comparison of the ONIOM-EC-RISM results with its extrapolation target.

A medium-sized biomolecular system suitable for this purpose was described by Grubmüller and coworkers in ref. [124, 125]. In this publication the authors calculated the pK_a values for four pentapeptides of the type Gly-X-Ala-X-Gly, or GXAXG for short, where X is either a glutamic acid or histidine residue, by *in-silico* titrations carried out using a constant-pH molecular dynamics approach.[124–129]

Furthermore, the authors conducted nuclear magnetic resonance (NMR) titrations to obtain experimental, site-specific reference pK_a values for the validation of their computational approach. The experimental shifts as a function of pH are shown in figure 6.1 and table 6.1. These experimental pK_a values and NMR chemical shifts will also serve as a reference for the ONIOM-EC-RISM predictions.

Although the NMR titration was used by the authors of the reference publication as a way of validating the results obtained from their constant-pH model, the EC-RISM model in general also allows the prediction of spectroscopic parameters. In order to test the capabilities of the ONIOM-EC-RISM model in this regard, an additional attempt is made to directly model the experimental NMR titration. For this purpose, a novel methodology is presented which allows to model the pH dependence of the predicted chemical shifts.

Of the four peptides presented by the authors, only the GEAEG peptide will be

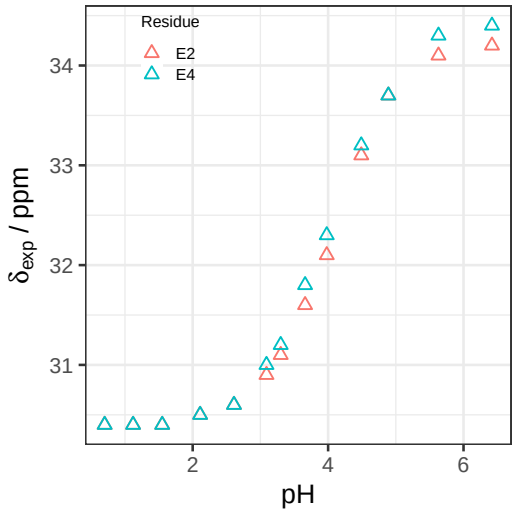

Figure 6.1.: Experimental chemical shifts of the glutamic acid C_γ atoms of the GEAEG pentapeptide as a function of pH, as measured by Grubmüller and coworkers.[124,125] Colours differentiate between residues E2 (red) and E4 (blue). The corresponding numerical values are shown in table 6.1.

Table 6.1.: Experimentally determined chemical shifts of the glutamic acid C_γ atoms of the GEAEG pentapeptide. The data were provided directly by the authors of the reference publication.[124,125]

pH	$\delta_{C_\gamma,exp}$ / ppm	
	E2	E4
0.7	30.4	30.4
1.12	30.4	30.4
1.55	30.4	30.4
2.11	30.5	30.5
2.61	30.6	30.6
3.09	30.9	31.0
3.3	31.1	31.2
3.66	31.6	31.8
3.98	32.1	32.3
4.49	33.1	33.2
4.89	33.7	33.7
5.63	34.1	34.3
6.42	34.2	34.4

studied here, although the methodology described below can also be easily applied to the remaining pentapeptides.

6.1. Modelling of the NMR titration experiment

6.1.1. MD-based conformer sampling

In their constant-pH MD simulations, Grubmüller and coworkers modelled the protonation of glutamic acid with a three-state model, where each of the two oxygen atoms of the carboxylic acid moiety can be protonated. The protonation of the terminal groups was fixed in the zwitterionic state, i.e. with a deprotonated C-terminus and a protonated N-terminus. In the constant-pH MD approach, the transitions between the three protonation states were modelled continuously using λ-dynamics.[129]

To model the same states with EC-RISM, a total of four microstates in three macrostates were considered, as shown in figure 6.2.

In the following, a four-digit code is used to indicate the protonation pattern of the GEAEG pentapeptide and thus to identify the individual microstates. A total of four protonation sites are used for this structure. These are the carboxylic acid groups of the C-terminus and the two glutamic acid side chains, as well as the amino group of the N-terminus. In the four-digit code, a one indicates a protonated state of the respective protonation site, while a zero indicates the corresponding conjugated base. The position of the protonation centres in the four-digit code corresponds to their position in the peptide, i.e. the first digit indicates the N-terminus, while the last digit indicates the protonation state of the C-terminus. Similar notations have been used in publications such as refs. [130, 131]

In contrast to the conformer generation workflow used for the EC-RISM predictions for the SAMPL5[23] and SAMPL6[24] challenges, an alternative approach is tested here where snapshots were taken directly from MD simulations of each microstate without further geometry optimisation.

As the snapshots are sampled directly from the respective trajectories, the microstate energies are calculated by the arithmetic mean

$$G_{it} = \frac{1}{N_c} \sum_{c}^{N_c} G^{\mathrm{corr}}_{\mathrm{sol},itc}, \tag{6.1}$$

where N_c is the total number of snapshots in the microstate t. This procedure is motivated by the fact that the set of snapshots sampled from the trajectory are already pre-weighted and therefore do not require further Boltzmann weighting. However, this introduces a sampling error due to the inherent inaccuracies of a force field based description of the pentapeptide energy landscape. The rest of the pK_a calculation procedure as described in section 5.2 remains unchanged.

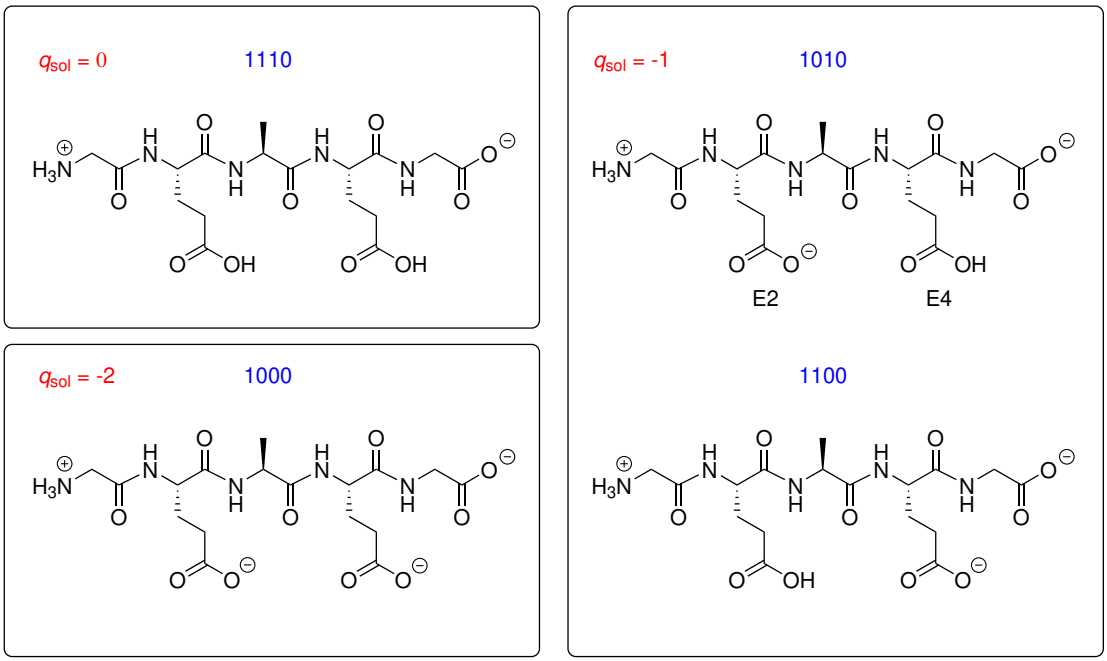

Figure 6.2.: GEAEG microstates considered for the ONIOM-EC-RISM calculations, labelled with a four-digit code indicating the protonation state of the four protonation centres, i.e. both glutamic acid side chains and both peptide chain termini. For example, the code "1010" indicates that the N-terminus and the C-terminal glutamic acid side chain (E4) are protonated, whereas the C-terminus and N-terminal glutamic acid side chain (E2) are deprotonated. The microstates are grouped by their corresponding macrostates.

6.1.2. Chemical shift calculation with ONIOM-EC-RISM

Until now, predictions of molecular properties have been made on the basis of free energies. However, the iterative process of an ONIOM(QM:SQM)-EC-RISM calculation also yields fully solvated wavefunctions, which can be used to calculate spectroscopic properties of the system under investigation.

Recently, Maste et al. demonstrated that the *hr*-EC-RISM approach can be used to accurately predict chemical shifts for the small model compounds, trimethylamine N-oxide and N-methylacetamide.[33] In this approach, after convergence of the main EC-RISM iterations, the (chemical) shieldings σ are obtained by evaluating the *real* system in the converged solvent point charge field at the *high* level of theory.

For the analogous calculation with ONIOM-EC-RISM, in principle the same approach has to be applied. Since the chemical shielding is a property localised at individual atoms, two cases have to be considered. One where the atom for which the shielding is

calculated is part of the *model* system $\mathcal{A}_\mathrm{m}$ and one where it is outside and the quantities of the *model* system calculations are not available, hence

$$\sigma_i^\mathrm{ONIOM} = \begin{cases} \Omega\left(\boldsymbol{\sigma}_i\right), & \text{if } i \in \mathcal{A}_\mathrm{m} \\ \sigma_i^\mathrm{lr}, & \text{otherwise.} \end{cases} \tag{6.2}$$

In the first case, the shielding of atom i is obtained from an ONIOM extrapolation of the shieldings obtained from the individual sub-calculations. In the second case only the *lr* result is available and the shieldings are calculated accordingly. It should be noted, however, that in most cases the choice of *model* system atoms will result in the shieldings of the atoms of interest being calculated according to the first case anyway. This calculation scheme can be applied to all ONIOM-EC-RISM schemes as the shielding is calculated in an already converged solvent point charge field.

In addition to these considerations arising purely from the partitioning of the system, further approximations have to be made due to the choice of codes for the implementation of the ONIOM-EC-RISM schemes presented here. As mentioned previously, in all cases the *high* level calculations are performed with ORCA, while the *low* level calculations are performed with EMPIRE. While the former offers a wide range of methods for calculating spectroscopic parameters, no such methods are available for EMPIRE. The choice of these codes for this implementation was initially made with the aim of extrapolating efficiently to the *hr* free energies, without any intention of predicting chemical shieldings. As a consequence, for this implementation equation 6.2 becomes

$$\sigma_i^\mathrm{ONIOM} = \sigma_i^\mathrm{hm}, \tag{6.3}$$

implicitly assuming

$$\sigma_i^\mathrm{lr} = \sigma_i^\mathrm{lm}, \tag{6.4}$$

with the additional constraint that the atom i must be part of the *model* system. The assumption is valid for partitions where the influence of the remaining system not included in the *model* system on the atom i is minimal. If the PFL is used, this additional approximation gives the same result as if the *low* level shieldings were available from EMPIRE.

From these shieldings the chemical shift δ_i can be calculated directly with

$$\delta_i = \sigma_{\mathrm{ref},i} - \sigma_i^\mathrm{ONIOM}, \tag{6.5}$$

given an appropriate reference $\sigma_{\mathrm{ref},i}$.

It should also be noted that the implementation used here and the resulting unavailability of the *low* level shieldings means that no ONIOM extrapolation is effectively performed. Since only the *hm* shieldings are available and are evaluated in the point charge field of the solvent, and in the case of the EE scheme in the field of the remaining atoms, this approach is similar to additive multiscale models. Again it should be emphasised that this is only an effect of the implementation. In the case where all shieldings are available a subtractive multiscale model is used.

6.1.3. Calculation of pH-dependent population weighted chemical shifts

In order to model the NMR titration experiment conducted by Grubmüller and coworkers,[124] it is necessary to introduce a pH-dependence and a population weight into the predicted chemical shielding. For this purpose, an approach to calculate these weights, first presented by Jochen Heil[17] and later applied in the context of the SAMPL6 challenge by Tielker et al.[24] is adapted here.

In this approach, the pH-dependent population

$$x_i(\mathrm{pH}) = \left(10^{-\mathrm{pH}}\right)^{-i} \prod_{j=0}^{i} 10^{-\mathrm{p}K_{\mathrm{a},j}} \left(1 + \sum_{k=1}^{n} \left(10^{-\mathrm{pH}}\right)^{-k} \prod_{l=1}^{k} 10^{-\mathrm{p}K_{\mathrm{a},l}}\right)^{-1} \tag{6.6}$$

of a given macrostate i can be calculated from the macroscopic $\mathrm{p}K_\mathrm{a}$ values of the chemical system. As before, $i = 0$ corresponds to the species with the highest degree of protonation considered, while n refers to the number of titratable sites.

The population of a given microstate within a macrostate is expressed as

$$x_{it|i} = \frac{1}{Z_i} \exp\left(-\beta G_{it}\right), \tag{6.7}$$

where Z_i is the macroscopic partition function as defined by

$$Z_i = \sum_t \exp\left(-\beta G_{it}\right). \tag{6.8}$$

The total pH-dependent microstate population is therefore obtained by multiplication

$$x_{it}(\mathrm{pH}) = x_{it|i}\, x_i(\mathrm{pH}), \tag{6.9}$$

which can be interpreted as the probability of finding a tautomer t at a given pH, due to the normalisations $\sum_i x_i = 1$ and $\sum_t x_{it|i} = 1$.[17,24]

These microstate populations can now be used to introduce the pH dependent weights into the calculated chemical shifts. Similar to the free energies from equation 6.1, the microstate chemical shifts $\delta_{a,tc}$ for an atom a are obtained from the snapshots sampled from the MD trajectory via the arithmetic mean

$$\delta_{a,t} = \frac{1}{N_c} \sum_c \delta_{a,tc}. \tag{6.10}$$

Using x_{it} as the weight function, the pH dependent shift of the respective atom can then be obtained by the weighted sum

$$\delta_a(\mathrm{pH}) = \sum_t \delta_{a,t}\, x_{it}(\mathrm{pH}). \tag{6.11}$$

The approach to modelling experimental NMR titrations can therefore be divided into two parts. First, the prediction of the pK_a values is carried out using the modified conformer sampling approach presented above. Secondly, the ONIOM-EC-RISM model can be used to additionally predict the chemical shifts of the microstates, which can then be population-weighted using the previously calculated pK_a values to model the pH-dependence observed in the experiment.

6.1.4. Calculation of site-specific titration curves

In addition to the previous approach of weighting the NMR shieldings with an appropriate population function, an approach presented by Ullmann in 2003[130] is used for the direct calculation of site-specific titration curves. Although Ullmann's method has been generalised to allow the study of polyprotic acids with N protons, the following discussion is restricted to the diprotic case, as a diprotic approximation is also used for the pentapeptide by fixing the protonation of the termini. In the following the four microstates of this diprotic model are abbreviated as 00, 10, 01 and 11. Note that this notation is virtually identical to the four-digit code used for the pentapeptides, as the protonation of the termini, i.e. the first and last digits, is fixed.

To model the experimental titration curves, microstate populations can be calculated from the microstate pK_a values. Although the population notation used in the previous section was useful for iterating over all relevant states, the limited number of microstates makes it possible to refer to them directly as $\langle(00)\rangle$, $\langle(10)\rangle$, $\langle(01)\rangle$ and $\langle(11)\rangle$, which improves readability. Given the microstate pK_a values pK_a^b defined for the transition from microstates a to b, the microstate probabilities can be calculated as

$$\langle(00)\rangle = \frac{1}{Z} \tag{6.12}$$

$$\langle(10)\rangle = \frac{1}{Z} 10^{\mathrm{p}K_{10}^{00} - \mathrm{pH}} \tag{6.13}$$

$$\langle(01)\rangle = \frac{1}{Z} 10^{\mathrm{p}K_{01}^{00} - \mathrm{pH}} \tag{6.14}$$

$$\langle(11)\rangle = \frac{1}{Z} 10^{\mathrm{p}K_{10}^{00} + \mathrm{p}K_{01}^{00} - W - 2\mathrm{pH}}. \tag{6.15}$$

The microstate pK_a values can also be used to calculate the partition function

$$Z = 1 + 10^{\mathrm{p}K_{10}^{00} - \mathrm{pH}} + 10^{\mathrm{p}K_{01}^{00} - \mathrm{pH}} + 10^{\mathrm{p}K_{10}^{00} + \mathrm{p}K_{01}^{00} - W - 2\mathrm{pH}}, \tag{6.16}$$

where

$$W = \mathrm{p}K_{10}^{00} - \mathrm{p}K_{11}^{01} = \mathrm{p}K_{01}^{00} - \mathrm{p}K_{11}^{10} \tag{6.17}$$

denotes the interaction energy between the two protonation sites. Positive values of W therefore indicate that the binding of a proton by one site disfavours the binding of a

proton at the other site. Conversely, negative values would be obtained for systems with cooperative binding.[130]

By summing over the states where a particular site is protonated, the site-specific titration curves for that site are obtained, i.e.

$$x_1 = \langle(10)\rangle + \langle(11)\rangle \tag{6.18}$$

and

$$x_2 = \langle(01)\rangle + \langle(11)\rangle\,. \tag{6.19}$$

The macroscopic titration curve X is obtained by summing over all microscopic probabilities multiplied by their respective total number of bound protons:

$$X = \langle(10)\rangle + \langle(01)\rangle + 2\,\langle(11)\rangle\,. \tag{6.20}$$

Using the relation $\mathrm{p}K_{11}^{01} + \mathrm{p}K_{01}^{00} = \mathrm{p}K_{11}^{10} + \mathrm{p}K_{10}^{00}$ it can be concluded that the titration curves of a diprotic acid are parameterised by three of the four microstate $\mathrm{p}K_\mathrm{a}$ values or, as used later in this work, by $\mathrm{p}K_{01}^{00}$, $\mathrm{p}K_{10}^{00}$ and W.[130]

6.2. Computational details

Starting from the snapshots sampled from the microstate trajectories, single point calculations were performed with the ONIOM-EC-RISM/B scheme with electronic embedding, using MP2/6-311+G** as the *high* level method and PM6 for the *low* level, as for the previous $\mathrm{p}K_\mathrm{a}$ predictions. The settings for the ONIOM-EC-RISM calculations were identical to those used for the previous calculations on the SAMPL6 dataset. The resulting free energies were corrected with the "/B@PFL (All)" PMV correction parameters. The $\mathrm{p}K_\mathrm{a}$ values were then obtained by applying the *hr*-$\mathrm{p}K_\mathrm{a}$ correction parameters (/B|B|M|A‖*hr*|A). The /B scheme was chosen due to the drastic speed-up measured in section 5.3.9 compared to the other ONIOM-EC-RISM schemes, given the large number of snapshots sampled and the associated computational cost. The correction parameters were chosen because they gave the best predictions for the /B scheme and the SAMPL6 dataset in the previous chapter. In addition, the /B@PFL $\mathrm{p}K_\mathrm{a}$ correction was tested in conjunction with the "/B@PFL (All)" PMV correction (/B|B|M|A‖A), as it gave the second best result of all /B models on the SAMPL6 dataset and an RMSE equal to the *hr* reference model. See table 5.26 for an overview of these two correction models in the context of SAMPL6. Please refer to tables 5.5 and 5.11 for the PMV and $\mathrm{p}K_\mathrm{a}$ correction parameters.

Apart from the faster calculation of the free energies, the use of the /B scheme offers additional computational advantages for the prediction of chemical shifts. Since the single point calculations were performed using MP2 as the *high* level of theory, it would theoretically be possible to obtain the chemical shieldings at the same level of theory

directly from the final iteration after convergence of the solvent environment in the main EC-RISM iterations, if the *high* level provides a suitable method. However, in the publication by Maste et al.[33] it was shown that the density functional OLYP[132] with the 6-311+G** basis set gives accurate chemical shift predictions for the *hr*-EC-RISM scheme. It is therefore desirable to apply the same level of theory to the chemical shift predictions with ONIOM-EC-RISM. As the calculation of the solvent environment is decoupled from the *model* system calculations and thus from the *hm*-NMR calculation, the shieldings at the OLYP level of theory can be obtained by simply re-evaluating the *hm* sub-calculation with this functional, thus eliminating the need to repeat the entire ONIOM-EC-RISM/B single point calculation. This drastically reduces the overall computational effort to obtain the free energy prediction for the pK_a prediction and the chemical shielding at their respective levels of theory.

All MD simulations of the microstates 1000, 1100, 1010, 1110 and an additional state to be introduced later, 100NHMe, were performed using a ported version of the Amber99SB[133] force field for GROMACS 2023.1[134–136] and a timestep of 2 fs. The SPC/E water model[120] was used in all cases. Peptide bonds were constrained using the LINCS algorithm,[137] while water bonds were constrained using the SETTLE algorithm.[138] In addition, a cubic simulation box with a side length of 6 nm and a NaCl concentration of 0.15 M was used. The simulation temperature of 300 K and the pressure of 1 bar were controlled using the Nosé-Hoover thermostat[139,140] and the Parrinello-Rahman barostat.[141,142] Electrostatic interactions were treated using the particle mesh Ewald algorithm[143,144] with a cut-off distance of 1 nm and a Fourier grid spacing of 0.12 nm. Similarly, a cut-off distance of 1 nm was used for the Lennard-Jones potential. The input structures were first minimised using the steepest descent method, followed by minimisation using the conjugate gradient method. From these structures a 10 ns NVT simulation was performed, followed by a NpT simulation of the same length to equilibrate the system. Production simulations were then performed for 400 ns for each microstate. Peptide snapshots were sampled from the resulting trajectory at a rate of 4 frames/ns. Hence, a total of 1601 frames were sampled from each micro state trajectory.

All MD simulations were performed by Lars Schumann. The MD simulation protocol given above as well as the snapshots used as input for the ONIOM-EC-RISM single point calculations were also provided by Lars Schumann.

As it has previously been demonstrated that the addition of explicit water molecules improves the accuracy of chemical shielding predictions with *hr*-EC-RISM,[33] additional frames were generated containing explicit solvent molecules from the MD simulation in a radius of 2.5 and 4.0 Å around the studied γ-carbon atoms of the glutamic acid side chains. In contrast to the previous calculations, these have been obtained with a reduced sampling rate of 2 ns^{-1} to counteract the increased computational cost due to the explicit water molecules. These additional samples were provided by Stefan Maste. To aid convergence, the single point calculations of these structures were performed by performing the initial *lr*-calculation not in vacuum but in a PCM solvent, as detailed in

section 4.6.

6.3. ONIOM partitioning

In order to reduce the overall computational cost, the GEAEG system was partitioned for the ONIOM-EC-RISM/B single point calculations. These partitions are shown in figure 6.3 and are used for both the pK_a and the chemical shift predictions.

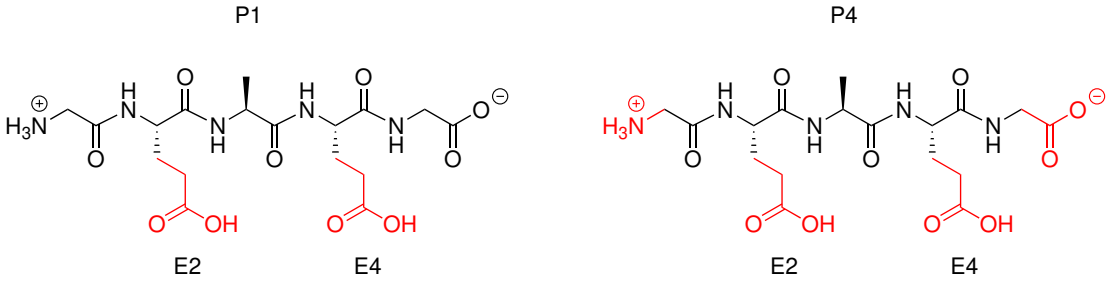

Figure 6.3.: ONIOM partitions applied to the GEAEG pentapeptide. Atoms marked in red are part of the *model* system. In all cases, open valencies were saturated with hydrogen link atoms. Both partitions result in *model* systems that include the side chains of both glutamic acid residues, as they contain the γ-carbon atoms that are studied in the NMR titration experiment, as well as both protonation centres and therefore need to be included in the model system. P1 includes only the glutamic acid side chains, while P4 also includes the charged parts of the terminal residues in order to include their polarising effect in the *high* level calculations.

P1 is the starting point for the analysis. It has a reduced *model* system containing only the glutamic acid side chains and therefore the two protonation centres of the system. No additional atoms are included in the *model* system to reduce the cost of the expensive *hm* sub-calculation. It was decided not to reduce the *model* system further, e.g. to the terminal carboxylic acid moieties, because the γ-carbon of the side chain was studied as part of the NMR titration experiment[124] and therefore needs to be modelled using the *high* level method.

P4, the second partitioning considered here, was created in the course of the analysis described in the following sections in order to investigate the influence of the charged peptide chain termini on the calculated properties. In addition to the part of the system already included in the first partitioning, P4 also includes the C-terminal carboxy moiety and the N-terminal amino group. The cuts have been chosen to include the adjacent methylene group in order to cut bonds that result in *model* system fragments that are expected to be less polarised due to the additional methyl group introduced.

The numbering of the partitions presented here simply reflects the order in which they were created. Partitions P2 and P3 were created as test cases that do not add to the

arguments made in this chapter and are therefore not presented.

As P1 only includes the glutamic acid side chains in the *model* system, the polarising effect of the charged terminal residues is modelled by EE-charges. In contrast, P4 also includes the termini and therefore allows the polarisation effect to be modelled at the *high* level of theory, i.e. MP2 for pK_a calculations and OLYP for chemical shieldings calculations.

6.4. Acidity constant predictions

6.4.1. Convergence analysis

As the conformer sampling approach employed here differs significantly from those previously used within the SAMPL challenges,[23,24,145] it is necessary to first validate it.

To do this, the convergence of the conformer ensemble is estimated by considering the convergence of the resulting pK_a values. For this analysis, the underlying microstate free energies are calculated by the cumulative mean, which calculates the average over the ensemble up to a given simulation time. Evaluating this cumulative mean over the full length of the trajectory gives a mean free energy as a function of simulation time.

To formally define this cumulative mean, the entire simulation interval is first discretised into the series $t_{\mathrm{sim},1}, \ldots, t_{\mathrm{sim},n}, \ldots, t_{\mathrm{sim},N_c}$ to reflect the N_c conformers sampled from the individual microstate simulations. The ONIOM-EC-RISM calculations result in the corresponding time series of conformer free energies $G_{it1}, \ldots, G_{itc}, \ldots, G_{itN_c}$, which can then be used to calculate the time series of cumulatively averaged microstate free energies with

$$G_{it,n} = \frac{1}{n} \sum_{c=1}^{n} G_{itc}. \tag{6.21}$$

As a result, the micro and macro state pK_a values are also a function of the simulation time and are used here to provide a first estimate of whether the sample size and simulation time are sufficiently large to allow converged pK_a values to be obtained.

In order to assess whether the overall computational cost can be reduced by reducing the conformer sampling rate of $4\,\mathrm{ns}^{-1}$, the set of conformers was resampled with reduced sampling rates of 2, 1 and $0.5\,\mathrm{ns}^{-1}$. Using this approach, both macroscopic and microscopic pK_a values, defined as the direct transition between two microstates in adjacent protonation states, were calculated. The latter are labelled with the microstates involved in the deprotonation reaction, e.g. "1110 - 1100" is the microstate pK_a for the transition of 1110 to its conjugated base 1100. The corresponding pK_a values calculated for all four sampling rates are given in table 6.2, while the convergence analysis obtained by the cumulative mean are shown in figures 6.4 and 6.5.

The statistical errors were obtained by calculating the errors from the standard deviations of the microstate free energies and propagating the errors to the macroscopic pK_a

values. The reported statistical errors thus represent the uncertainties introduced by the MD-based sampling and neglect the additional contributions introduced by the four empirical correction parameters. The errors should therefore be considered as a lower bound on the true statistical error.

For the resampling analysis, the highest sampling rate is used as a reference for the highest feasible computational effort. By comparing this reference with the lower sampling rates, it is clear that only the $2\,\mathrm{ns}^{-1}$ sampling rate is able to reproduce the more expensive pK_a predictions. For both sampling rates pK_a values are obtained that differ only in the second decimal place. In contrast, lower sampling rates give pK_a values that differ more significantly. Since the two highest sampling rates both give almost identical results, it can be concluded that sampling above $4\,\mathrm{ns}^{-1}$ is unlikely to change the resulting pK_a values significantly, while only increasing the computational effort required. Furthermore, to increase the overall computational efficiency of the sampling approach, a sampling rate of $2\,\mathrm{ns}^{-1}$ would have been sufficient for this system.

The reference sampling rate of $4\,\mathrm{ns}^{-1}$ in combination with the *hr*-MP2 pK_a correction gives macroscopic pK_a values of -0.66 and 4.34 for P1 and 0.33 and 3.40 for P4. In both cases these values are almost identical to one of the microstates. In the case of the higher pK_a value this is the transition "1010 - 1000", while the lower one is identical to "1110 - 1010". This is a first indication that the transition from state 1110 to 1000 occurs via 1010. It is therefore likely that the side chain carboxylic acid of residue E2 is deprotonated before E4.

For the second pK_a correction, /B@PFL, the same observations can be made, i.e. the identity between the macro pK_a values and some micro pK_a values. However, when compared to the *hr*-MP2 correction, different numerical values are obtained. Here macroscopic pK_a values of 0.26 and 4.18 for P1 and 1.03 and 3.44 are predicted.

The statistical errors reported in table 6.2 are significantly smaller than the associated pK_a values and, as expected, decrease as the sampling rate increases. However, it should be emphasised that this can only be taken to mean that the underlying sample sizes are sufficiently large to yield convergent pK_a values if the underlying conformer samples are uncorrelated. Due to the exploratory nature of this pentapeptide study, no effort has been made to investigate possible correlation effects of the samples on the target quantities, although this is planned for further studies based on this work. Nevertheless, it can be assumed that the large time steps between frames, even at the highest sampling rate, are sufficient to achieve an uncorrelated ensemble.

The pK_a values discussed previously were obtained by averaging over the entire length of the simulation. Additional information can be obtained by considering the convergence of the time series of pK_a values calculated from the cumulative mean in equation 6.21. The corresponding results for the *hr*-MP2 and /B@PFL pK_a correction are shown in figure 6.4 and 6.5.

In addition to the assessment of the statistical error of the pK_a values and the purely graphical evaluation of the convergence behaviour, the numerical values of the time-

Table 6.2.: Predicted micro- and macrostate pK_a values and statistical errors for the GEAEG pentapeptides. The results are given for the two partitions P1 and P4 used for the single point calculations, the sampling rate used to obtain the conformers from the MD trajectory, as well as the applied pK_a correction. "*hr*-MP2" refers to the correction model with the ID /B|B|B|A‖*hr*|A, while "/B@PFL" is the corresponding model with the ID /B|B|B|A‖A. See the text for an explanation of the pK_a IDs and tables 5.5 and 5.11 for the corresponding PMV and pK_a parameters.

pK_a corr.	pK_a ID	Part.	p$K_{a,corr}$			
			0.5 / ns	1 / ns	2 / ns	4 / ns
hr-MP2	1010 - 1000	P1	4.77±0.55	4.73±0.38	4.33±0.27	4.34±0.19
		P4	3.72±0.54	3.65±0.36	3.34±0.25	3.39±0.18
	1100 - 1000	P1	2.18±0.57	2.10±0.39	2.09±0.28	2.03±0.20
		P4	2.40±0.56	2.38±0.38	2.41±0.26	2.35±0.19
	1110 - 1010	P1	-0.50±0.54	-0.97±0.37	-0.62±0.26	-0.66±0.18
		P4	0.31±0.52	0.12±0.37	0.38±0.26	0.34±0.18
	1110 - 1100	P1	2.09±0.56	1.66±0.38	1.62±0.27	1.64±0.19
		P4	1.62±0.55	1.39±0.38	1.31±0.27	1.38±0.19
	Macro pK_a 1	P1	-0.50±0.32	-0.97±0.22	-0.62±0.15	-0.66±0.11
		P4	0.30±0.31	0.11±0.22	0.36±0.15	0.33±0.11
	Macro pK_a 2	P1	4.77±0.32	4.73±0.23	4.33±0.16	4.34±0.11
		P4	3.72±0.32	3.66±0.22	3.36±0.15	3.40±0.11
/B@PFL	1010 - 1000	P1	4.52±0.43	4.49±0.30	4.17±0.21	4.18±0.15
		P4	3.69±0.42	3.64±0.29	3.40±0.20	3.43±0.14
	1100 - 1000	P1	2.49±0.45	2.42±0.31	2.41±0.22	2.37±0.16
		P4	2.66±0.44	2.64±0.30	2.66±0.21	2.62±0.15
	1110 - 1010	P1	0.38±0.42	0.02±0.29	0.29±0.20	0.26±0.14
		P4	1.02±0.41	0.87±0.29	1.07±0.20	1.04±0.14
	1110 - 1100	P1	2.41±0.44	2.08±0.30	2.04±0.21	2.07±0.15
		P4	2.05±0.43	1.86±0.30	1.80±0.21	1.86±0.15
	Macro pK_a 1	P1	0.38±0.25	0.02±0.17	0.29±0.12	0.26±0.08
		P4	1.01±0.24	0.86±0.17	1.06±0.12	1.03±0.08
	Macro pK_a 2	P1	4.52±0.25	4.49±0.18	4.17±0.12	4.18±0.09
		P4	3.70±0.25	3.65±0.17	3.41±0.12	3.44±0.08

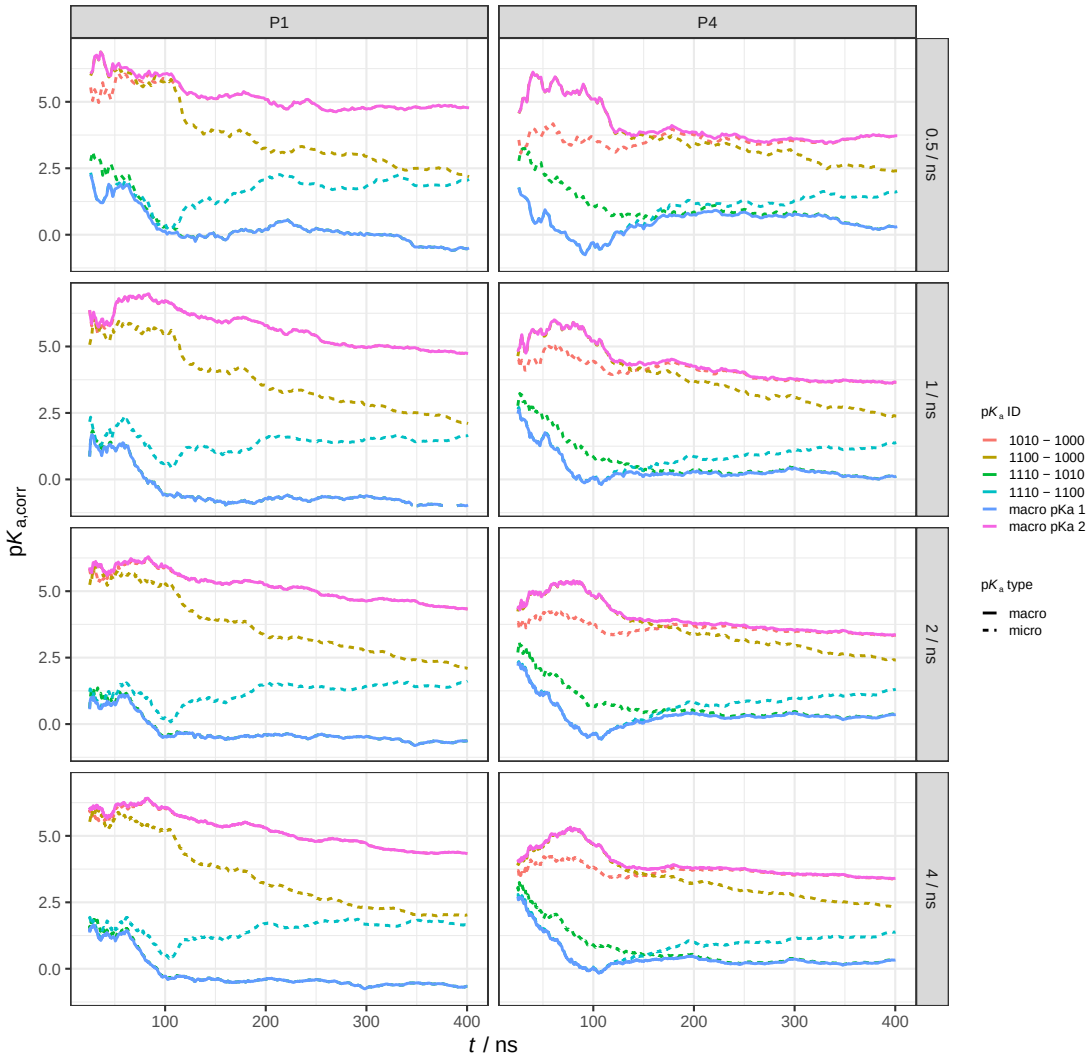

Figure 6.4.: Predicted pK_a values for the GEAEG pentapeptide as a function of simulation time. The pK_a values were obtained using the *hr*-pK_a correction, as explained in the main text. The underlying microstate free energies were calculated by cumulative averaging over all frames sampled from the trajectory up to the given simulation time t. The respective columns represent the applied ONIOM partitioning, while the rows indicate the sampling rate. The colours differentiate between pK_a IDs. See the text for an explanation of these IDs.

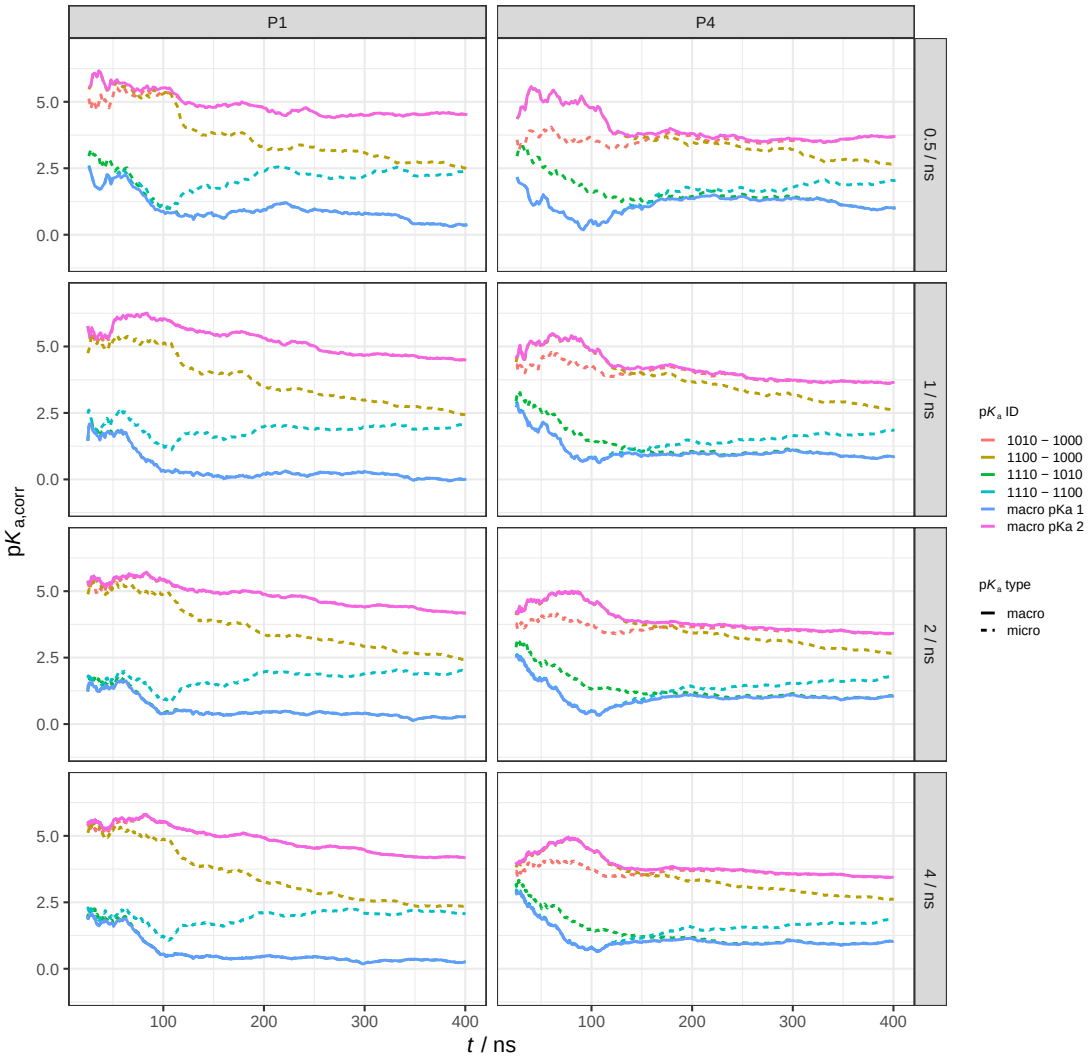

Figure 6.5.: Predicted pK_a values for the GEAEG pentapeptide as a function of simulation time. The pK_a values were obtained using the /B@PFL-pK_a correction, as explained in the main text. The underlying microstate free energies were calculated by cumulative averaging over all frames sampled from the trajectory up to the given simulation time t. The respective columns represent the applied ONIOM partitioning, while the rows indicate the sampling rate. The colours differentiate between pK_a IDs. See the text for an explanation of these IDs.

dependent pK_a changes for the highest sampling rate are shown in table 6.3. The values for all other frame rates are shown in tables 102 and 103 in the appendix. To analyse the convergence behaviour, the total simulation time was divided into sub-intervals and the sum of the changes in the cumulative pK_a value of two consecutive time points was calculated. This is mathematically identical to the difference between the left and right cumulative pK_a of the interval. It is expected that when the MD simulation has reached convergence, the pK_a value will cease to change as the number of frames over which it is averaged increases, thus reaching a plateau in the graphical plots. However, it should be emphasised that this is a purely qualitative measure to assess the drift of the cumulative pK_a value.

Table 6.3.: Change in pK_a for the given simulation time intervals in nanoseconds for the $4\,\mathrm{ns}^{-1}$ sampling rate and the two pK_a corrections considered in this chapter. The numerical values were calculated as the difference between the right and left values p$K_{a,corr,r}$ and p$K_{a,corr,l}$ of the given intervals. This is equal to the sum of all pK_a changes from one frame to the following and is used to estimate convergence of the pK_a for the complete set of MD simulation frames.

pK_a corr.	Part.	pK_a ID	p$K_{a,corr,r}$ − p$K_{a,corr,l}$						
			(50,100]	(100,150]	(150,200]	(200,250]	(250,300]	(300,350]	(350,400]
hr-MP2	P1	Macro pK_a 1	-1.38	-0.14	0.02	-0.02	-0.32	0.20	-0.13
		Macro pK_a 2	0.02	-0.69	-0.09	-0.51	-0.07	-0.34	-0.03
		1010 - 1000	0.10	-0.66	-0.09	-0.51	-0.07	-0.34	-0.03
		1100 - 1000	-0.44	-1.38	-0.64	-0.55	-0.35	-0.28	0.01
		1110 - 1010	-1.47	-0.16	0.02	-0.02	-0.32	0.20	-0.13
		1110 - 1100	-0.93	0.56	0.56	0.02	-0.04	0.14	-0.17
	P4	Macro pK_a 1	-1.37	0.23	0.17	-0.24	0.15	-0.21	0.17
		Macro pK_a 2	0.03	-0.94	0.03	-0.01	-0.22	-0.02	-0.14
		1010 - 1000	-0.30	-0.32	0.21	0.03	-0.22	-0.00	-0.14
		1100 - 1000	0.07	-1.11	-0.32	-0.25	-0.22	-0.26	-0.16
		1110 - 1010	-1.05	-0.38	-0.01	-0.27	0.15	-0.22	0.17
		1110 - 1100	-1.42	0.41	0.52	0.00	0.15	0.03	0.20
/B@PFL	P1	Macro pK_a 1	-1.09	-0.11	0.01	-0.01	-0.25	0.16	-0.10
		Macro pK_a 2	0.01	-0.54	-0.07	-0.40	-0.06	-0.27	-0.03
		1010 - 1000	0.08	-0.52	-0.07	-0.40	-0.06	-0.27	-0.03
		1100 - 1000	-0.34	-1.08	-0.50	-0.43	-0.28	-0.22	0.01
		1110 - 1010	-1.15	-0.12	0.01	-0.01	-0.25	0.16	-0.10
		1110 - 1100	-0.73	0.44	0.44	0.02	-0.03	0.11	-0.14
	P4	Macro pK_a 1	-1.08	0.18	0.14	-0.19	0.12	-0.16	0.13
		Macro pK_a 2	0.02	-0.73	0.02	-0.01	-0.17	-0.02	-0.11
		1010 - 1000	-0.23	-0.25	0.16	0.02	-0.17	-0.00	-0.11
		1100 - 1000	0.05	-0.87	-0.25	-0.20	-0.17	-0.20	-0.13
		1110 - 1010	-0.82	-0.30	-0.01	-0.21	0.12	-0.17	0.13
		1110 - 1100	-1.11	0.32	0.41	0.00	0.11	0.03	0.15

From figures 6.4 and 6.5 it can be seen that the convergence behaviour for both pK_a

corrections is almost identical. A comparison of the values in table 6.3 leads to the same conclusion. The following discussion is therefore limited to the *hr*-pK_a correction, but can be transferred to the /B@PFL correction.

For P1 and all sampling rates, averaging over the increasing simulation length initially leads to a rapid decrease in the lower macroscopic pK_a value "macro pK_a 1" up to about $t_{sim} = 100\,ns$. Subsequently, the change is small, and in absolute terms it is less than 0.15 pK_a units for the last part of the simulation.

The second macroscopic pK_a value initially decreases continuously. However, in the last simulation interval of 350 to 400 ns the changes are small, with a value of -0.03 pK_a units. The picture is similar for the microscopic pK_a. Again only small changes can be observed in the last time interval. The largest absolute change can be seen for the transition 1110 - 1100 with a value of -0.17 pK_a units.

P4 shows a similar trend. Again the first macroscopic pK_a decreases rapidly and changes only slightly after $t_{sim} = 100\,ns$. The second macroscopic pK_a shows a similar behaviour. However, in contrast to P1, P4 shows larger changes in the last time interval for both macroscopic pK_a values, although these are also small with a maximum absolute value of 0.17.

As with P1, the cumulative pK_a value changes most in the last time interval for the 1110 - 1100 transition with a value of 0.20.

The samples taken from the highest sampling rate of $4\,ns^{-1}$ means that for each of the 4 microstates a number of 1600 single point calculations have to be performed for the given simulation time. Although the use of the fast ONIOM-EC-RISM/B method instead of *hr*-EC-RISM makes it possible to perform this large number of calculations at all, the calculation represents a large computational effort. Therefore in addition to the low statistical error of the predicted pK_a values and the small observed change, it was decided not to extend the simulation time or to use a higher sampling rate to include more conformers in the calculated ensemble.

6.4.2. Calculation of pH-dependent populations

In order to further investigate the pentapeptide system and as a prerequisite for the subsequent modelling of the NMR titration experiment, pH-dependent population curves were calculated according to the theory outlined in section 6.1.3. The results from the highest sampling rate of $4\,ns^{-1}$ are shown in figure 6.6.

Both pK_a corrections again give similar results, the only difference being that the distance between the intersection of the 1010 population curve with 1100 and 1000 is smaller for /B@PFL than for *hr*-MP2. This is due to the smaller difference between the microscopic pK_a values as shown in table 6.2.

As expected, in the pH ranges below the smallest pK_a values, the neutrally charged macrostate shows the highest population, while at pH values above the largest pK_a value, the double negatively charged macrostate is populated. In these pH ranges only

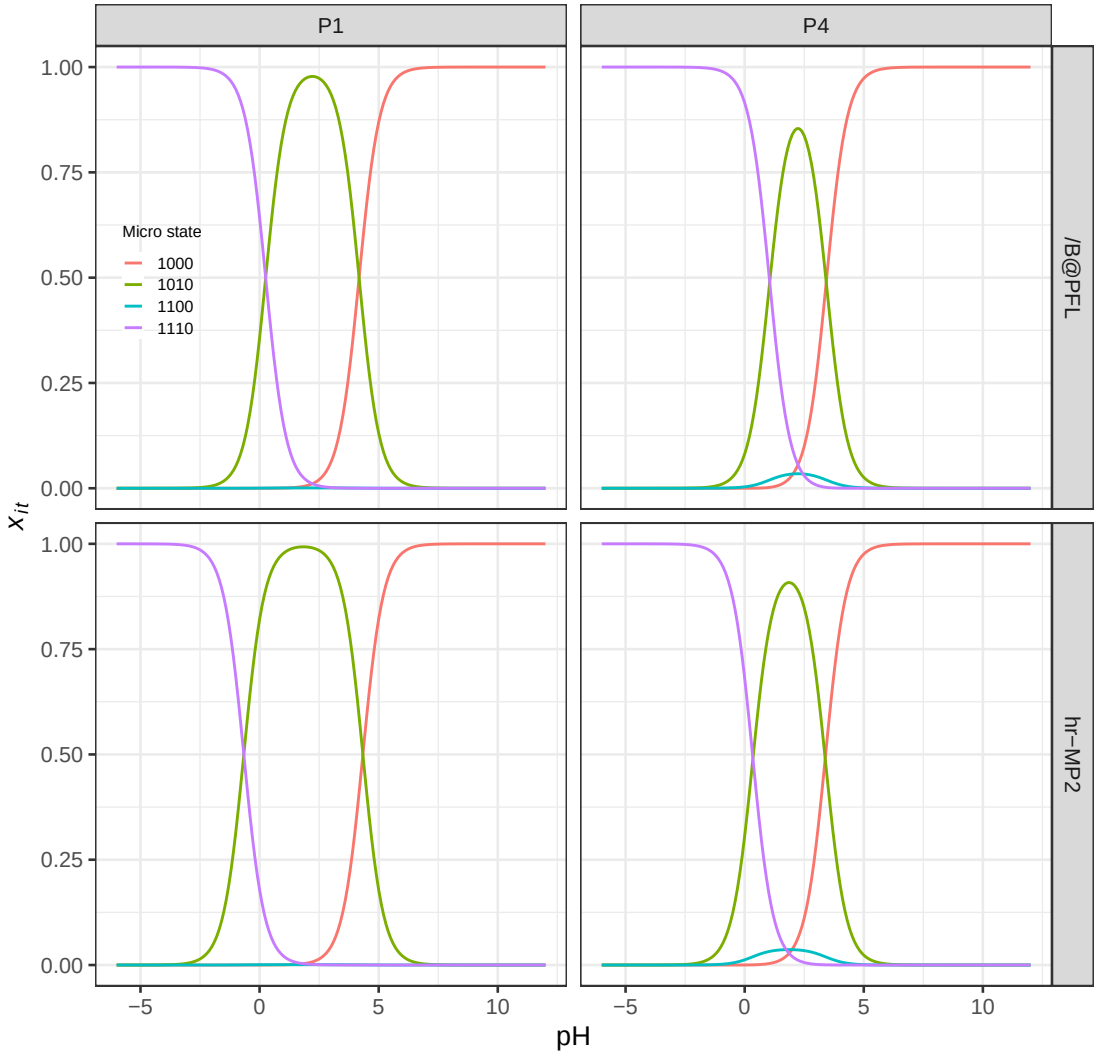

Figure 6.6.: Predicted pH-dependent microstate population curves calculated by equation 6.9 for the GEAEG pentapeptide from the $4\,\mathrm{ns}^{-1}$ sampling rate. Plots in the same row show population curves obtained using the same pK_a correction parameters, while columns indicate whether partition P1 or P4 was applied. Colour codes differentiate between the microstates "1000" (red), "1010" (green), "1100" (blue), "1100" (purple).

the 1110 and 1000 states are populated as no other microstates were modelled within these macrostates.

In contrast, for the single negatively-charged macrostate, which is populated in the pH range between the other two macrostates, two underlying microstates are available. Of these, only 1010 is significantly populated for both partitions.

Assuming that the macroscopic pK_a values are accurately predicted by EC-RISM, this suggests that as pH increases, the side chain of E2 is deprotonated first, followed by E4. The validity of this assumption will be further investigated in the next section by comparison with reference values.

6.4.3. Comparison to site-specific reference values

The reference publication by Grubmüller and coworkers[124] reports pK_a values from two sources for the GEAEG pentapeptide. The first pK_a values were obtained from λ-dynamics simulations and the second, experimental values, were obtained from NMR titrations. In both cases the reported pK_a values are site-specific, i.e. they are specifically assigned to the carboxy group of either glutamic acid residues E2 or E4.

This of course differs from the pK_a calculation methodology presented in this work and in previous publications using EC-RISM.[23,24,122,145] The macroscopic pK_a values presented in table 6.2 are a function of the difference between the ensemble averages of two adjacent macrostates. Although the underlying microstates represent a transition between two specific tautomers, the macroscopic averages no longer contain information about the specific protons involved in the deprotonation process, i.e. one can say that a proton has been removed from the system, but not which one, as the macrostate no longer differentiates between two microstates that share the same number of protons.

In contrast to micro-state pK_a values, the pK_a values obtained from NMR titrations and λ-dynamics simulations represent a process where the investigated site is deprotonated, while the influence of all other protonation sites and their changes is summed. However, microstate pK_a values describe the transition between two microstates in two adjacent macrostates. Consequently, in the case of the investigated GEAEG pentapeptide and the diprotic approximation with four states considered here, only a specific group is deprotonated, while the protonation of the remaining sites is fixed.[131] Therefore, neither macro nor micro pK_a provide the same localised information as the site-specific pK_a reported in the reference paper.

While only the macroscopic and microscopic pK_a values are required for the subsequent modelling of pH-dependent populations and chemical shifts, the calculation of site-specific pK_a values allows additional insight into the pentapeptide system and allows direct comparison with reference values without the need to explicitly model the NMR experiment. However, the development of the methodological framework required to calculate these site-specific pK_a values based on free energy predictions with EC-RISM is beyond the scope of this work, but may be revisited in future publications.

In this respect, a series of publications by Ullmann and coworkers[130,131,146,147] dealing with the calculation of site-specific pK_as from electrostatic models may provide a good starting point. Here the authors present two conceptually similar measures, Tanford-Roxby[148] and Henderson-Hasselbalch pK_a values, which converge to nearly the same value for weak interactions between protonation sites. The authors state that these pK_a measures are motivated by the shortcomings of the half-protonation point p$K_{1/2}$, a concept recently rediscovered by Fraczkiewicz and coworkers,[149] which cannot adequately determine pK_a values from flatter protonation or irregularly shaped titration curves.[131]

However, in ref. [131] Bombarda and Ullmann show that the highest and lowest microscopic pK_a values associated with the deprotonation of a site are lower and upper bounds of the site-specific pK_a value of that site. For the GEAEG system this means that the site-specific pK_a value of E2 that would be obtained from EC-RISM lies between the microscopic pK_a values for the "1100 - 1000" and "1110 - 1010" transitions. Therefore, for P1, the highest sampling rate and the *hr*-MP2 pK_a correction, a site specific value between -0.66 and 2.03 would be obtained. For P4 the corresponding value would be between 0.34 and 2.35. From the experiment and the λ-dynamic simulations, site-specific pK_a values of 4.06 and 3.81 were obtained for E2, respectively. Thus, both reference values are well outside the range that would be predicted by EC-RISM.

Using the same reasoning, the site specific pK_a for E4 would lie between the micro pK_a values of "1110 - 1100" and "1010 - 1000", giving a range of 1.64 to 4.34 for P1. For P4 values between 1.38 and 3.39 can be expected. Here the two reference values of 4.04 and 4.09 fall within the range predicted by EC-RISM for P1 and outside for P4, although the difference between the upper limit and the reference values is less than 0.8 pK_a units.

These observations for E2 and E4 are also true for the /B@PFL-pK_a correction.

It is therefore likely that the acidity of the E2 side chain is drastically overestimated by the EC-RISM model, while the acidity of the other glutamic acid residue E4 is likely to be modelled more accurately. As the local acidity of E2 was estimated using the two microstate pK_as as lower and upper bounds, it can also be assumed that one or both of the microstate transitions "1100 - 1000" and "1110 - 1010" are not modelled accurately. If it is indeed the case that the micro pK_a obtained for the transition "1110 - 1010" is too small, then the smaller macro pK_a will also be underestimated as it is identical to the micro pK_a.

Ultimately, this means that the transition between the macrostates with charges of 0 and -1 would occur at higher pH values than shown in figure 6.6.

Although no attempt will be made in this work to calculate site-specific pK_a values directly on the basis of the above publications, Ullmann's publication from 2003[130] provides a methodology for calculating site-specific titration curves (see section 6.1.4), which will be used here to allow a qualitative comparison with the results of the NMR titration experiment shown in figure 6.1. This is mainly used as an illustrative example to

highlight the microscopic pK_a values that would need to be obtained from the ONIOM-EC-RISM model to reproduce the shape of the sigmoidal titration curves presented in the reference publication by Grubmüller and coworkers.[124,125]

As outlined in section 6.1.4, the titration curves x_1, x_2 and X are defined by three parameters, pK_{01}^{00}, pK_{10}^{00} and W. Therefore, the effect of these parameters on the shape of the titration curves will first be qualitatively elucidated. For this purpose, four sets of microscopic pK_a values and the interaction energy W are given in table 6.4 and the resulting titration curves are shown in figure 6.7.

Table 6.4.: Definition of the set of pK_a values used as an illustrative example for the calculation of site-specific titration curves of a diprotic acid from figure 6.7, based on Ullmann's approach.[130] Here pK_a^b denotes the microscopic pK_a value for the transition from state a to b and W is the interaction energy as defined by Ullmann (equation 6.17).

ID	pK_{11}^{01}	pK_{11}^{10}	pK_{01}^{00}	pK_{10}^{00}	W
$(4.0, 3.0, W = 4)$	-1.0	0.0	4.0	3.0	4.0
$(4.0, 4.0, W = 4)$	0.0	0.0	4.0	4.0	4.0
$(4.0, 4.0, W = 2)$	2.0	2.0	4.0	4.0	2.0
$(4.0, 4.0, W = 0)$	4.0	4.0	4.0	4.0	0.0

The first set of parameters $(4.0, 3.0, W = 4)$ defines a system where the acidity of one protonation site is slightly more acidic than the other and there is a strong interaction of $W = 4$ between the sites. As a consequence, there is a clear stepwise deprotonation visible in the macroscopic titration curve X, while the two site-specific curves x_1 and x_2 differ significantly. With increasing pH, the average protonation of the site described by x_1 decreases rapidly to an approximate value of 0.1 before reaching a plateau. A second smaller step is seen between pH 4 and 5, after which the site is completely deprotonated. The other site shows the opposite behaviour. First a small decrease in the average degree of protonation x_2 is observed, before a large decrease is observed between pH 4 and 5.

In contrast, the NMR titration experiment presented by Grubmüller and coworkers (figure 6.1) suggests that both sites are deprotonated almost simultaneously, resulting in two almost identical sigmoidal curves. It is therefore obvious that the first set of model micro pK_a values are not suitable to reproduce the experimental curves.

The second set of parameters $(4.0, 4.0, W = 4)$ describes a system where both sites have the same acidity, as postulated by Grubmüller and coworkers for the GEAEG pentapeptide.[124,125] The interaction of these sites remains unchanged compared to the previous set ($W = 4$). Consequently, both site-specific titration curves are identical. Both show a stepwise partial deprotonation, first to a plateau with a value of 0.5 and then to complete deprotonation of the respective site, which still differs from the sigmoidal curve obtained from the NMR titration experiment.

For the last two sets of parameters, the pK_a values are identical to the previous set,

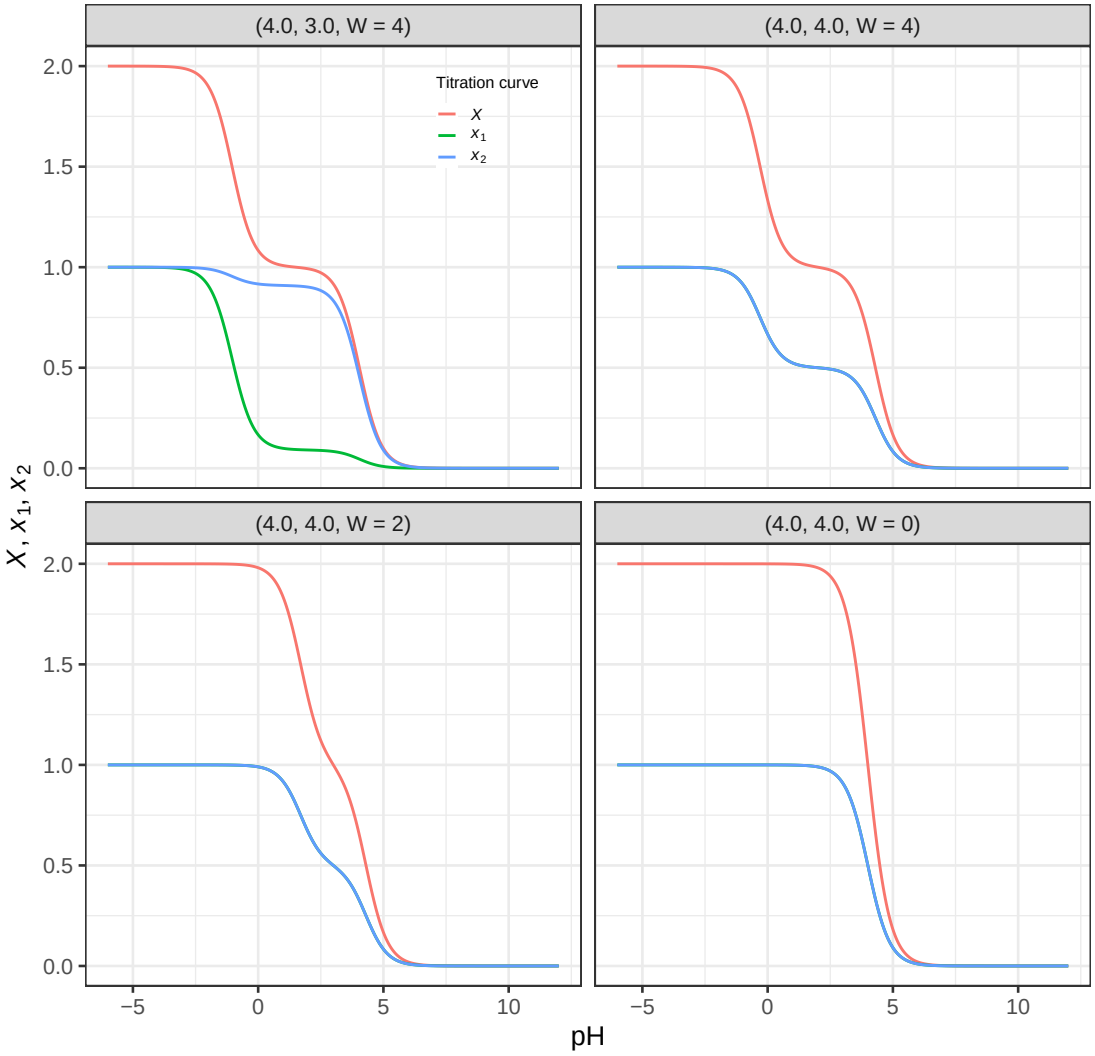

Figure 6.7.: Illustrative example of site-specific titration curves for the two-state approximation of GEAEG pentapeptide obtained from the set of microstate pK_a values shown in table 6.4, based on Ullmann's approach.[130] The colours distinguish the macroscopic titration curve X (red) and the two site-specific titration curves x_1 (green) and x_2 (blue).

Table 6.5.: The set of microscopic pK_a values used to calculate the site-specific titration curves from figure 6.8, based on Ullmann's approach for diprotic acids.[130] These were obtained from the ONIOM-EC-RISM/B model using the highest sampling rate of $4.0\,\mathrm{ns}^{-1}$ (table 6.2). Here pK_a^b denotes the microscopic pK_a value for the transition from state a to b and W is the interaction energy as defined by Ullmann (equation 6.17).

pK_a corr.	Part.	pK_{1110}^{1010}	pK_{1110}^{1100}	pK_{1010}^{1000}	pK_{1100}^{1000}	W
hr-MP2	P1	-0.66	1.64	4.34	2.03	2.70
	P4	0.34	1.38	3.39	2.35	2.00
/B@PFL	P1	0.26	2.07	4.18	2.37	2.11
	P4	1.04	1.86	3.43	2.62	1.57

but W decreases to values of 2 and 0, modelling systems where the interaction between sites is less pronounced $(4.0, 4.0, W = 2)$ or absent $(4.0, 4.0, W = 0)$. From figure 6.7 it is clear that decreasing W results in a less pronounced intermediate protonation step, which is completely missing for $W = 0$, giving a sigmoidal curve. This can be explained by the fact that neglecting the interaction of the system in the model effectively splits the system into two independent monoprotic acids.

It can therefore be assumed that, given a diprotic approximation of the pentapeptide, the sigmoidal titration curves of the experiment can be reproduced theoretically if a model predicts identical microscopic pK_a values for the transitions "1010 - 1000" and "1100 - 1000" and only a weak interaction between the protonation sites E2 and E4.

The titration curves obtained from the microscopic pK_a values calculated at the ONIOM-EC-RISM/B level on structures sampled at a rate of $4.0\,\mathrm{ns}^{-1}$ are shown in figure 6.8. The corresponding microscopic pK_a values and the interaction energy W are given in table 6.5.

For both pK_a corrections it can be seen that larger interaction energies W are predicted for partition P1 compared to P4. At the same time, the /B@PFL-pK_a correction gives smaller values for W than the *hr*-pK_a correction. Consequently, the model using the /B@PFL-pK_a correction and P4 shows the smallest W of 1.57. At the same time, this model shows the smallest difference between pK_{1010}^{1000} and pK_{1100}^{1000}, giving results that are close to the above conditions that must be met to reproduce the experimental titration curve. However, it is clear from figure 6.8 that none of the models are able to reproduce the experiment accurately.

As the model used here was previously shown to accurately predict pK_a values for the SAMPL6 dataset, it is likely that the MD-based sampling approach is responsible for the deviations from the experimental reference, as it is the only methodological difference to the previous prediction approach. This may be due to the inherent inaccuracies associated with the force field basis description of the pentapeptide conformational energy

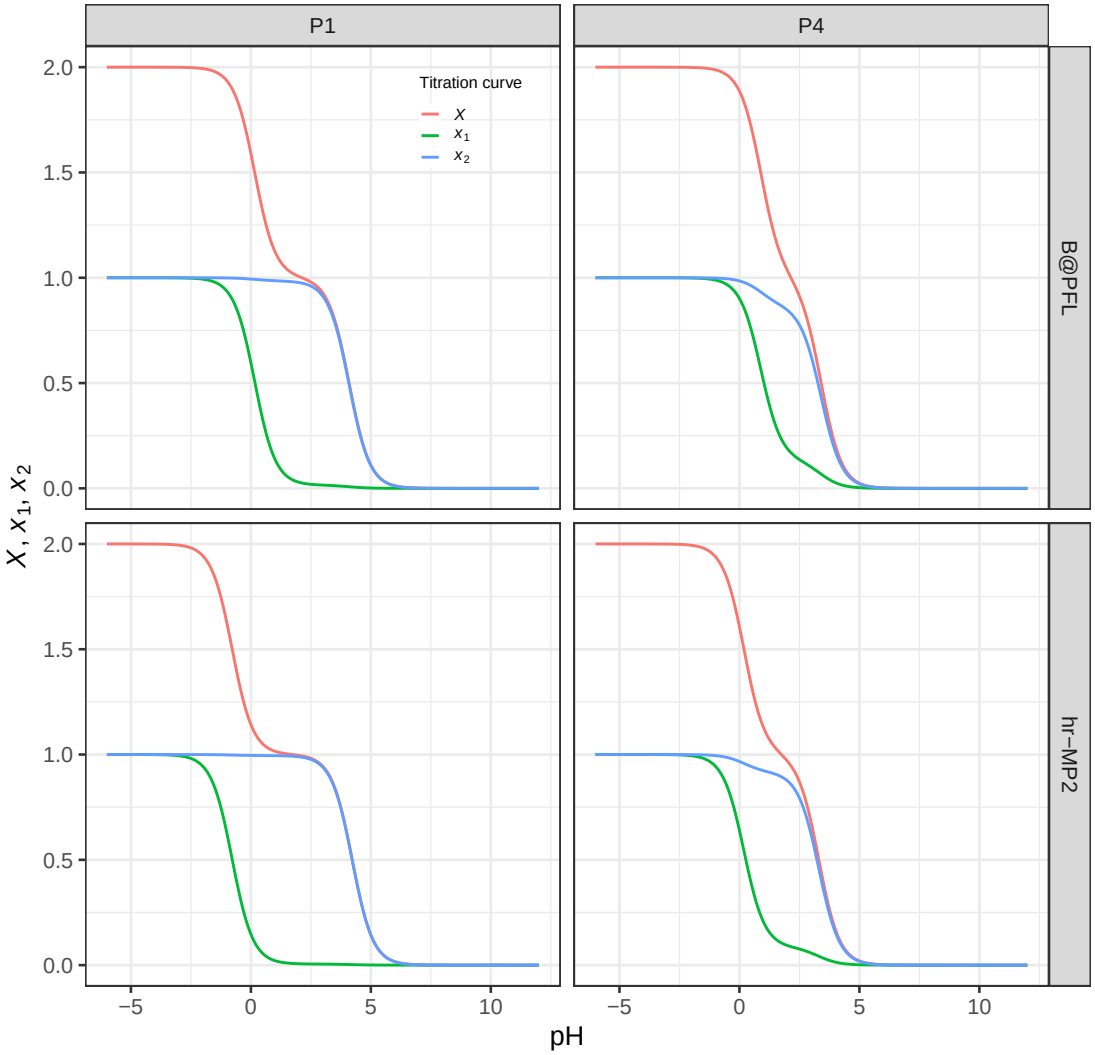

Figure 6.8.: Site-specific titration curves for the two-state approximation of GEAEG pentapeptide, based on Ullmanns approach.[130] Here the microstate pK_a values obtained from the ONIOM-EC-RISM/B calculations at the 4.0 ns^{-1} sampling rate (tables 6.2 and 6.5) were used to calculate the titration curves. The rows indicate the applied pK_a correction, while the columns indicate the partitions. Colours distinguish the macroscopic titration curve X (red) and the two site-specific titration curves x_1 (green) and x_2 (blue).

landscape. In contrast, for the SAMPL6 dataset the conformers were optimised at the B3LYP-PCM level and therefore at a higher level of theory.

6.5. Chemical shift predictions

6.5.1. Initial model evaluation

Using the populations calculated above as weights for the chemical shifts, it is possible to model the NMR titration curve directly, as already described in section 6.1.3. To do this, the microscopic shifts must first be obtained by averaging over the respective conformer ensembles. The resulting quantities for the C_γ atom of the E2 and E4 side chains are given in table 6.6. In the following all chemical shifts are obtained from the highest sampling rate of $4\,\mathrm{ns}^{-1}$. The chemical shifts resulting from the NMR titration experiment are given in table 6.1 and figure 6.1.

Table 6.6.: Averaged microstate chemical shifts for the C_γ atom of residues E2 and E4 of the GEAEG pentapeptide. All quantities were obtained from the $4\,\mathrm{ns}^{-1}$ sampling rate. The NMR titration experiment (table 6.1) gives shifts of 30.4 ppm for both residues at pH 0.7. At pH 6.42 shifts of 34.2 and 34.4 ppm are obtained for E2 and E4 respectively. These two pH values are the two end points of the experiment.

Part.	Residue	$\delta_{C_\gamma,t}$ / ppm			
		1000	1010	1100	1110
P1	E2	36.97±0.06	37.11±0.06	35.86±0.06	36.06±0.06
P1	E4	37.32±0.07	34.81±0.06	37.29±0.06	36.27±0.07
P4	E2	37.84±0.06	38.02±0.06	36.97±0.06	37.18±0.06
P4	E4	36.93±0.07	34.31±0.06	36.81±0.07	35.90±0.07

For the C_γ atom of E2 a general trend is visible for both partitions. Starting from microstate 1000, protonation of the E2 side chain leads to a significant decrease in the shift of -1.11 ppm for P1 and -0.87 ppm for P4, whereas protonation of the E4 side chain increases the shift, although the effect is minimal. This suggests that, due to the proximity, the protonation of the E2 carboxyl group influences the shift more than the protonation state of E4.

Protonation of both residues again leads to a decrease in the shift. In fact, the E2 shift of the doubly protonated state 1110 can be artificially constructed by considering the isolated protonation effects on the unprotonated state 1000. More precisely, adding the shift changes from 1000 to 1010 and from 1100 to 1000, i.e.

$$\delta_{C_\gamma,1110} = \delta_{C_\gamma,1000} + \Delta_{1000\to1010}\,\delta_{C_\gamma} + \Delta_{1000\to1100}\,\delta_{C_\gamma} \tag{6.22}$$

with

$$\Delta_{1000\rightarrow1010}\,\delta_{C_\gamma} = \delta_{C_\gamma,1010} - \delta_{C_\gamma,1000} \tag{6.23}$$

and

$$\Delta_{1000\rightarrow1100}\,\delta_{C_\gamma} = \delta_{C_\gamma,1100} - \delta_{C_\gamma,1000}, \tag{6.24}$$

gives chemical shifts for E2 of 36.00 ppm for P1 and 37.15 ppm for P4, which in both cases are almost identical to the explicitly calculated values from table 6.6.

The C_γ atom of E4 shows the opposite behaviour. Protonation of the E2 residue leads to a small increase in the shift, whereas protonation of E4 leads to a significant decrease. It is worth noting that the shift decrease for the transition from 1000 to 1010 is more than twice as large as the corresponding shift for the transition from 1000 to 1100 for E2. The C_γ atom of E4 therefore seems to be more influenced by the protonation of its adjacent carboxyl group than E2.

The application of equation 6.22, this time to artificially construct the E4 micro state shifts of 1110, yields 34.78 ppm for P1 and 34.19 ppm for P4. In contrast to the outcomes obtained for E2, the explicitly calculated shifts are significantly underestimated by this approach.

Figure 6.9 shows the NMR titration curve resulting from the populations shown in figure 6.6 and the chemical shifts from table 6.6, in addition to the experimental values from the reference publication.[124] As noted above, the results for the two pK_a corrections are very similar due to the similarity in their pH-dependent populations. Therefore, the following discussion is limited to the results obtained from the *hr*-pK_a correction, but can be transferred to the /B@PFL correction. The corresponding /B@PFL based results are shown in figures 8 and 9 in the appendix.

From figure 6.6 it is clear that neither the predictions for E2 nor E4 can reproduce the NMR titration experiment. On the one hand, the absolute predicted values are smaller than the reference. On the other hand, the relative difference between the maximum and minimum observed values is also too small compared to the experiment.

While the general sigmoidal shape of the experimental titration curve can be reproduced in the experimental pH range for E4, i.e. an increase in the observed shift with increasing pH, the other glutamic acid residue E2 shows an inverted response. In the lower experimental pH ranges, EC-RISM predicts that the microstate 1010 is populated. With increasing pH, 1000 becomes increasingly populated. As the predicted chemical shift of 1010 is slightly larger than that of 1000, the trend observed in the experiment is reversed for the predicted titration curve. This applies to both partitions. It is therefore necessary to improve the predictions of these initial models for both residues.

6.5.2. Evaluating the influence of explicit local water molecules

Recently, Maste et al. showed that the addition of explicit water molecules to the statistical solvent background significantly improves the predictions made with *hr*-EC-RISM.[33]

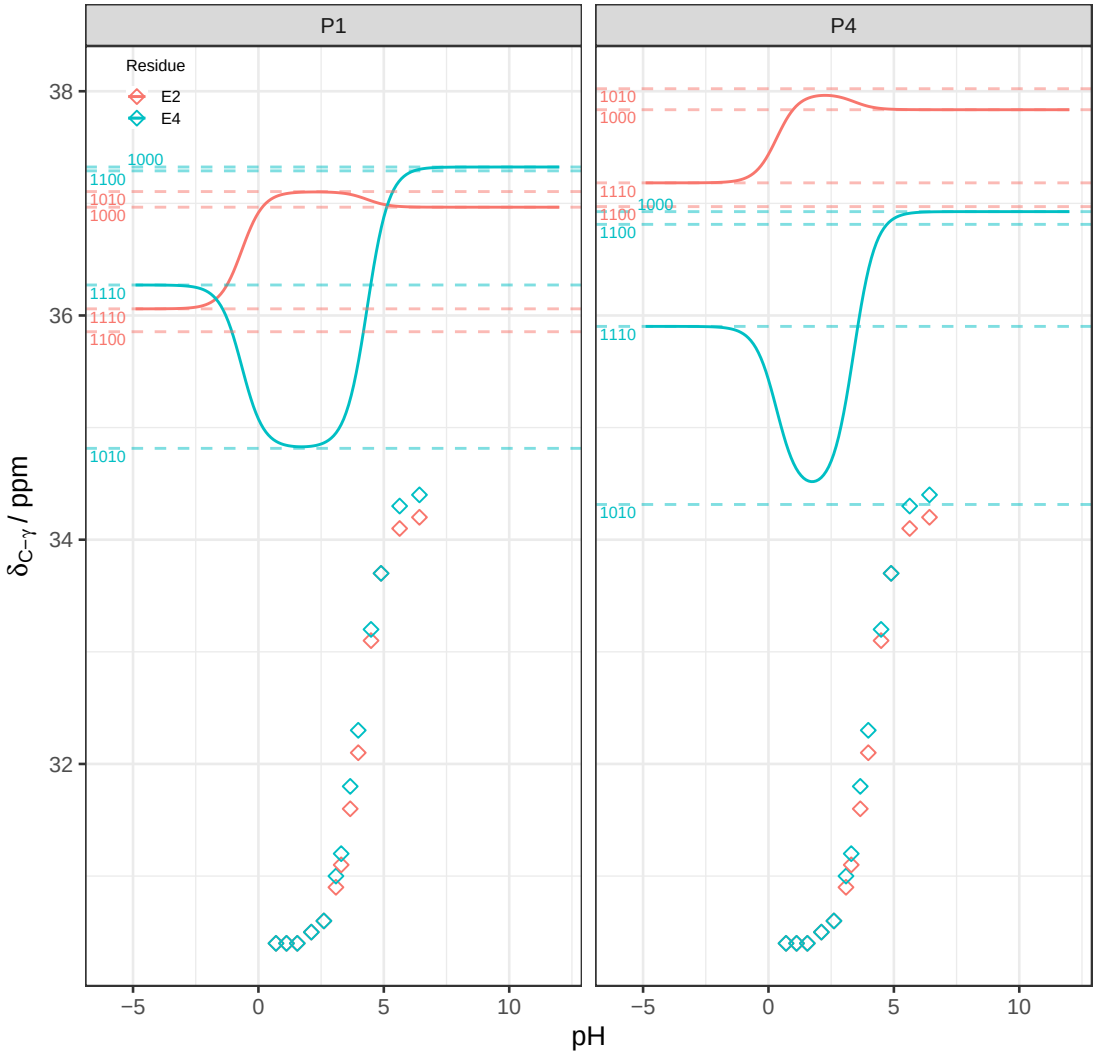

Figure 6.9.: Predicted pH-dependent chemical shift curves (equation 6.11) for the C_γ atoms of the GEAEG pentapeptide. The underlying populations were obtained using the *hr*-pK_a correction. Experimental chemical shifts are shown as diamonds. The colours distinguish between the results obtained for either residue E2 (red) or E4 (blue). Dashed horizontal lines show the microstate shifts from table 6.6.

Therefore, the same approach is tested here to improve the previously presented ONIOM-EC-RISM results.

The averaged microstate shifts based on the application of the 2.5 and 4.0 Å explicit solvation shells are shown in table 6.6 in comparison with the previous results without explicit solvation.

Table 6.7.: Averaged microstate chemical shifts for the C_γ atom of residues E2 and E4 of the GEAEG pentapeptide from calculations without and with explicit water shells of 2.5 and 4.0 Å. Samples without explicit water were obtained with a sampling rate of $4\,ns^{-1}$, while samples with explicit water were obtained with a reduced sampling rate of $2\,ns^{-1}$. Also shown are results for the 111NHMe state, which shows methylamination of the C-terminus.

Solvation	Part.	Residue	$\delta_{C_\gamma,t}$ / ppm				
			1000	1010	1100	1110	111NHMe
EC-RISM	P1	E2	36.97±0.06	37.11±0.06	35.86±0.06	36.06±0.06	34.68±0.06
	P1	E4	37.32±0.07	34.81±0.06	37.29±0.06	36.27±0.07	35.93±0.06
	P4	E2	37.84±0.06	38.02±0.06	36.97±0.06	37.18±0.06	-
	P4	E4	36.93±0.07	34.31±0.06	36.81±0.07	35.90±0.07	-
EC-RISM + 2.5 Å	P1	E2	35.40±0.20	35.42±0.18	34.35±0.18	34.72±0.17	-
	P1	E4	35.51±0.18	33.25±0.18	35.94±0.17	34.81±0.19	-
	P4	E2	36.31±0.21	36.38±0.18	35.39±0.18	35.86±0.17	-
	P4	E4	35.09±0.19	32.73±0.18	35.39±0.18	34.43±0.19	-
EC-RISM + 4.0 Å	P1	E2	33.50±0.20	33.46±0.18	32.73±0.18	33.05±0.16	-
	P1	E4	33.58±0.18	31.65±0.17	33.95±0.17	33.05±0.18	-
	P4	E2	34.52±0.20	34.49±0.19	33.77±0.19	34.19±0.16	-
	P4	E4	33.10±0.19	31.10±0.17	33.37±0.18	32.66±0.18	-

The results shown in this table are also visually summarised in figure 6.10 in comparison with the experimental results. For the latter, the end points of the titration curve, i.e. pH 0.7 and 6.4, are plotted as points. All other experimental points lie on the line connecting these two points in figure 6.10.

Similarly, the averaged microstate shifts obtained from the EC-RISM calculations with and without explicit solvation are shown as points. Since the weight function used to calculate the pH-dependent chemical shifts (equation 6.11) is a convex combination of these microstate shifts, all points of the predicted titration curve must also lie on the line connecting their points in figure 6.10. Therefore, in order to accurately reproduce the NMR titration experiment, it is essential that the experimental values fall within the range of the predicted microstate shifts. In other words, the range of the intervals shown in figure 6.10 can be used as a first indicator of whether the models tested are able to accurately reproduce the NMR titration experiment at all.

In both table 6.6 and figure 6.10, additional results are given for a 111NHMe state that shows methylamination of the C-terminus. These results will be discussed in the

following section, but in order to avoid repetition of almost identical tables and figures in the main part of this work, they are already shown alongside the results discussed in this section. For the sake of completeness, the figure omitting the 111NHMe state is shown in figure 7 in the appendix.

The addition of explicit solvent molecules clearly influences the calculated microstate shifts. While the calculations without additional water overestimate the experimentally observed range of shifts and consequently cannot reproduce the titration curve, the addition of the 2.5 Å water shell significantly reduces the shift for all microstates towards the range observed in the experiment. The addition of more water molecules for the 4 Å water shell further reduces the calculated shifts. This effect is observed for both partitions and residues.

In the case of P1, the 4 Å water shell shifts the microstates for both E2 and E4 into the experimental range. For P4 this is only true for E4, while for E2 the microstates 1000 and 1010 are outside the experimental range.

In addition to the trend towards smaller chemical shifts, the addition of explicit water molecules results in a reduction of the range observed for the microstate shifts. For example, the calculations without explicit water molecules yield a difference between the maximum and minimum shift of 1.25 ppm for P1 and E2. This range is reduced to a value of 1.07 ppm when the 2.5 Å water shell is added to the solute system. Adding more water molecules through the larger 4 Å solvation shell further reduces the range to 0.77 ppm.

A similar trend is seen for E4 and P1, although the range initially increases as the first explicit solvation shell is added. The addition of further solvent molecules then reduces the shift range again.

These trends for E2 and E4 can also be observed for the second partition P4. However, it can be observed that the shifts obtained for E2 are generally higher than those of P1. At the same time, the E4 shifts obtained from P4 are smaller than those of P1. As the *model* system generated by P4 contains both termini in addition to the *model* system atoms from P1, this shift must be attributed to their explicit modelling.

It can therefore be assumed that the chemical shifts of the C_γ atoms are influenced by the interaction between the adjacent side chains, carboxyl groups and the termini of the peptide. Thus, in addition to the microstates already studied, it would be useful to study other states in which the protonation state of the termini has been modified. These will be discussed in the next section.

6.5.3. Approximation of missing microstates and force field limitations

The previous calculations were carried out with a protonated N-terminus and a deprotonated C-terminus, resulting in a charged terminus in both cases. In order to further investigate the interaction of the termini with the glutamic acid side chains, and at the same time to keep the additional computational effort resulting from the modelling of

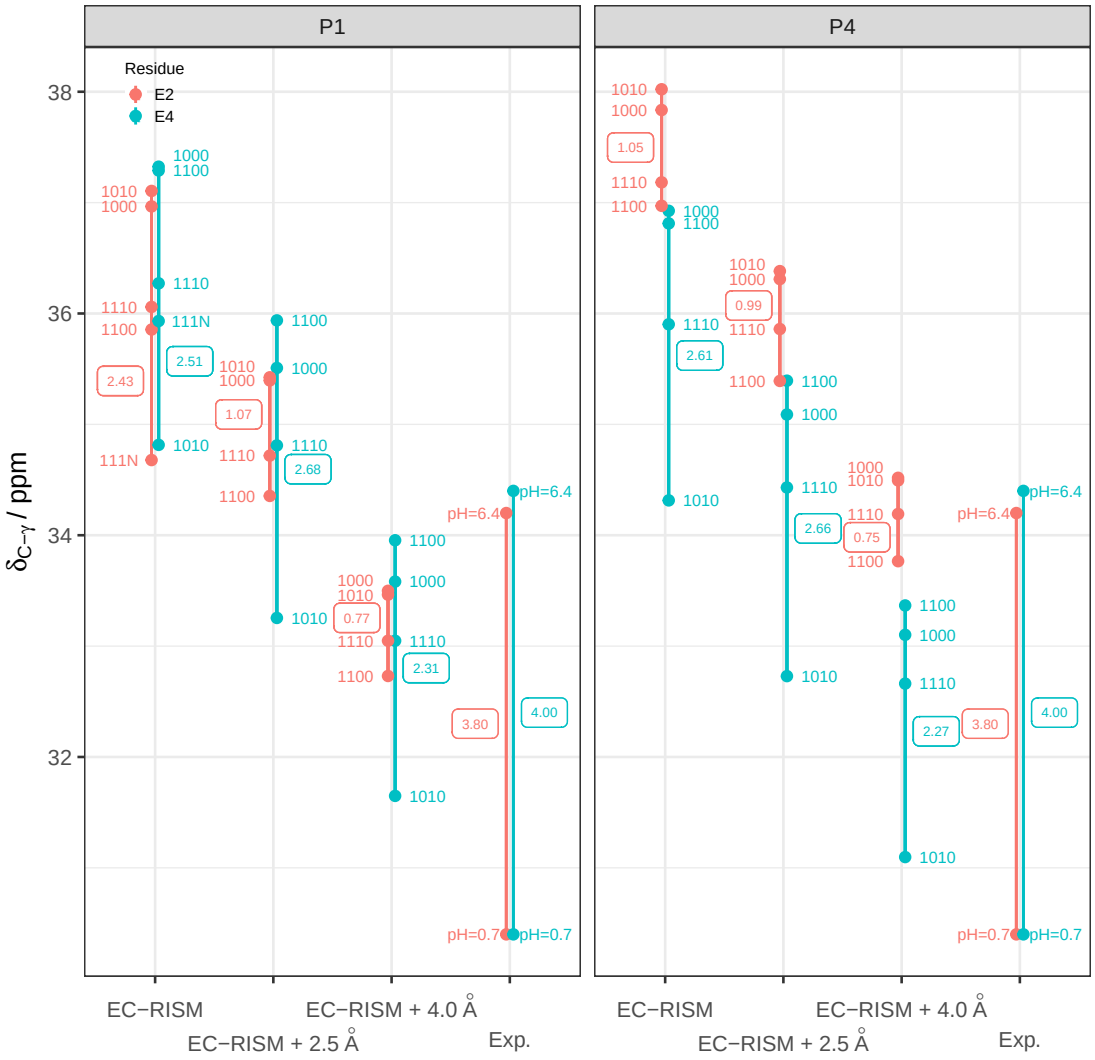

Figure 6.10.: Visualisation of the calculated chemical shifts with and without explicit water molecules in comparison with the experimental results. The points represent the averaged microstate chemical shifts or the end points of the NMR titration experiment, i.e. the measured shifts at pH 0.7 or 6.4. For the experiment, the lines represent the values that can be observed within the measured pH range. The lines connecting the calculated values represent the values that can be obtained for the pH-dependent weighted chemical shifts and can therefore be used to make an initial estimate of whether the method can reproduce the experiment. The values given in the boxes indicate the length of the lines and thus the difference between the maximum and minimum microstate shifts for the respective method or the range of shifts observed in the experiment. This range for the EC-RISM calculations without explicit water and without the additional 111NHMe state, abbreviated as 111N in this figure, is 1.25 ppm for E2, as also shown in figure 7 in the appendix.

the additional microstates low, it was decided to model only the protonation change at the C-terminus.

It can be expected that, due to the chemical similarity of the carboxyl group of the glutamic acid side chains and the C-terminus, the protonation transition also takes place at similar pH values and is likely to be coupled. This would also lead to a change in the chemical environment and presumably a change in the chemical shifts of the side chains. As this protonation transition is expected to occur in a similar pH range as the protonation state transition of the side chains, it is also expected to have a greater influence on the side chains compared to the N-terminus at this pH range.

In order to evaluate the influence of the protonated C-terminus, the complete set of states 1001, 1011, 1101 and 1111 would have to be modelled. As a first test, it was decided to model only one of these states, 1111.

Thus far, the states have been simulated using ff99SB,[133] which is the same force field employed in the reference publication by Grubmüller and coworkers.[124] While the carboxyl groups of the glutamic acid side chain can be parameterised in both their ionised and neutral protonation states, this force field only provides the zwitterionic parameterisation, in which the N-terminus is protonated and positively charged, while the C-terminus is deprotonated and negatively charged.

This, of course, greatly complicates the modelling of the protonated C-terminus. In order to model the uncharged state, the carboxyl group of the C-terminus was converted into its methylamide, which ultimately results in an uncharged terminus. This state is abbreviated as 111NHMe, and the corresponding micro state shifts are illustrated in table 6.7 and figure 6.10. The respective test calculations were carried out for the EC-RISM calculations without explicit water and partition P1.

The presented data demonstrate that the modification of the C-terminus has a pronounced effect on the E2 carboxyl function. A significant decrease in chemical shifts compared to the other micro states is observed, resulting in an increase of the total shift range from 1.25 to 2.43 ppm, as shown in figure 6.10.

As 111NHMe attempts to approximate the fully protonated state 1111, the effect of protonation of the C-terminus can be assessed by direct comparison with the adjacent state 1110. With a downshift of approximately -1.4 ppm for E2 and -0.3 ppm for E4, this comparison further demonstrates that the change in the C-terminal protonation state has a greater effect on the E2 C_γ atom than on the corresponding E4 atom.

It is important to note, however, that these results are derived by introducing a significant approximation to the chemical environment of the C-terminus. In order for the pronounced effect of the C-terminus on the E2 side chain to be valid, it is necessary to assume that the conversion of the C-terminus into its methylamide is a reasonable approximation for the protonation of the C-terminus. Although the approximation neutralises the charge of the carboxylate group as required, the introduction of a large functional group at a position where there is only one proton in 1111 may have unintended consequences. This approximation may introduce additional interactions, for

example, hydrogen bonds between the N-H group of the C-terminal amide and the side chains' carboxylate groups. This could potentially result in altered conformations sampled from the trajectory in comparison with the approximated fully protonated state 1111.

Nevertheless, this potential conformational effect cannot be readily evaluated without explicitly simulating this reference state. Consequently, one viable solution is to transition to a force field that provides the missing parameters for a protonated C-terminus. A suitable choice would be the ff14SB force field, developed by Maier et al.[150] This offers both protonated and deprotonated states for the C-terminus, and is also similar in parameterisation to the ff99SB force field, which has been applied in this work, as well as the reference study by Grubmüller and coworkers.

However, the application of a force field that can directly model a protonated C-terminus would also require the repetition of all previous molecular dynamics simulations and EC-RISM calculations for consistency. In light of the considerable computational cost involved, it was decided that it is preferable to accept the inherent uncertainties associated with approximating the protonated C-terminus by the NHMe group, as this approach is simply more resource-efficient.

So far, the fully protonated state 1111 has been used via its proxy 111NHMe to gain first insights into the protonation effect of the C-terminus on the glutamic acid side chains. However, it can be argued that this state is unlikely to be populated in the experimental pH range of 0.7 to 6.4. By calculating the pH-dependent population curves (equation 6.9 and figure 6.6), it was found that of the modelled states, only 1010 is likely to be highly populated at low pH. Decreasing the pH leads to protonation of the other E2 residue. A fully protonated state, regardless of the order in which the sites are protonated, would therefore be expected to occur at even lower pH values, probably well outside the experimental range.

It is therefore clear that if one wishes to investigate the influence of additional C-terminally protonated microstates on the lower experimental pH range of the NMR titration curve, these states must be sampled from the single negatively charged macrostate and not from the fully protonated macrostate. This means that the side chain proton of states 1010 and 1100 must be shifted to the C-terminus, resulting in the new microstate 1001.

Due to the force field limitations discussed above, this state would have to be approximated by the proxy state 100NHMe, as was done for 1111 and 111NHMe. However, it was decided that no further simulations should be performed and that these simulations will be part of future work based on this thesis.

Hence another way of artificially constructing the chemical shift of the 1001 microstate must be found. It has been shown previously that using equation 6.22 it is possible to accurately extrapolate the C_γ shift of E2 in 1111 by adding the individual protonation side chain effects to the unprotonated state 1000. Similarly, by calculating the C-terminal

protonation effect with

$$\Delta_{1110\rightarrow 111\mathrm{NHMe}}\,\delta_{\mathrm{C}_\gamma} = \delta_{\mathrm{C}_\gamma,111\mathrm{NHMe}} - \delta_{\mathrm{C}_\gamma,1110}, \tag{6.25}$$

the required state 100NHMe is approximated by

$$\delta_{\mathrm{C}_\gamma,100\mathrm{NHMe}} = \delta_{\mathrm{C}_\gamma,1000} + \Delta_{1110\rightarrow 111\mathrm{NHMe}}\,\delta_{\mathrm{C}_\gamma}, \tag{6.26}$$

which is then used as a proxy for the corresponding C-terminally protonated micro state 1001. The resulting shifts are shown in table 6.8.

Table 6.8.: Chemical shifts for the C_γ atom of residues E2 and E4 of the extrapolated 100NHMe micro state.

Solvation	Part.	Residue	$\delta_{\mathrm{C}_\gamma,100\mathrm{NHMe}}$ / ppm
EC-RISM	P1	E2	35.58
	P1	E4	36.98
	P4	E2	36.45
	P4	E4	36.59
EC-RISM + 2.5 Å	P1	E2	34.02
	P1	E4	35.17
	P4	E2	34.93
	P4	E4	34.75
EC-RISM + 4.0 Å	P1	E2	32.12
	P1	E4	33.24
	P4	E2	33.14
	P4	E4	32.76

Note that the protonation effect was estimated for the EC-RISM calculations without explicit water molecules and P1 partitioning, and was subsequently used to obtain all the results shown in table 6.8. As before, the 111NHMe state is assumed to be a reasonable approximation of the 1111 state. It is now also assumed that the resulting protonation effect can be transferred to the results for the explicit calculations and P4.

Incorporating the 1001 state into the NMR titration curve in a meaningful way is another challenge that requires some approximations. As there is no microstate population for 1001, no new population curve can be generated to weight the pH-dependent shifts. Instead, the shift of state 1010 is completely replaced by that of the extrapolated state 1001 and weighted with the previous population curve from figure 6.6.

However, here it must be also assumed that the population curve remains unchanged by the addition of the 1001 microstate. More specifically, the transition likely to correspond to the deprotonation of the E4 side chain is reassigned to the deprotonation

of the C-terminus, assuming that the associated macro pK_a and the population of the microstate remain unchanged.

This procedure was carried out for the EC-RISM results with the 4.0 Å explicit solvation shell, as this was the only solvation model that had previously shown results in the range of the experimental results in figure 6.10. The resulting NMR tritration curves, with the 1010 state replaced by the approximated 1001 state, are shown in figure 6.11 alongside the unmodified curves.

These unmodified curves confirm the results already obtained from the evaluation of the total shift range in figure 6.10. As the total microstate shift range for E2 and E4 was found to be smaller than the respective experimental shifts, the weighted pH-dependent shifts are also too small.

The unmodified E2 titration curve obtained from P1 shows a clear deviation from the experimental reference curve. For this residue, the 1010 microstate shift populated at the lower experimental pH range gives a shift only slightly smaller than the 1000 shift populated at higher pH values. Consequently, an almost constant weighted shift is predicted within the experimental range, whereas a range of 3.8 ppm is observed in the experiment.

This is also true for P4, although here the microstate shift of 1110 starts to become relevant at the lowest experimental pH values, as larger values are predicted for the lower of the two pK_a values, thus slightly increasing the overall shift range. However, this effect is too small to reproduce the experimental range.

For E4 the shift for 1010 is significantly smaller than that for 1000, resulting in a much larger shift range and results closer to the experiment than those obtained for E4. However, from the visualisation in figure 6.11 it is again clear that the minimum of the titration curve is too large, while its maximum is too small. This is true for both partitions.

This is possible for two reasons. First, the populated microstates 1010 and 1000 are in fact the states that are populated within the experimental pH range, but their predicted shifts are inaccurate, resulting in an underpredicted shift range. Second, it is possible that other states besides 1010 and 1000 are significantly populated, so that they would influence the predicted weighted shift and their exclusion leads to the observed inaccuracies. In particular, if two or more microstates contribute significantly to the weight function at a given pH, a combination of both effects is also possible.

However, the investigation of the first effect would require significant changes to the underlying model and possible conformer sampling, e.g. by switching to a more suitable force field as described above, which is beyond the scope of this work. This is also true for the second effect, as the consistent inclusion of additional microstates in the population curve would also require their conformers to be sampled by additional MD simulations and subsequent EC-RISM calculations. Here, only a partial aspect of the second effect can be investigated by replacing the 1010 microstate with the approximated 100NHMe state and reusing the pH-dependent weights without updating the associated

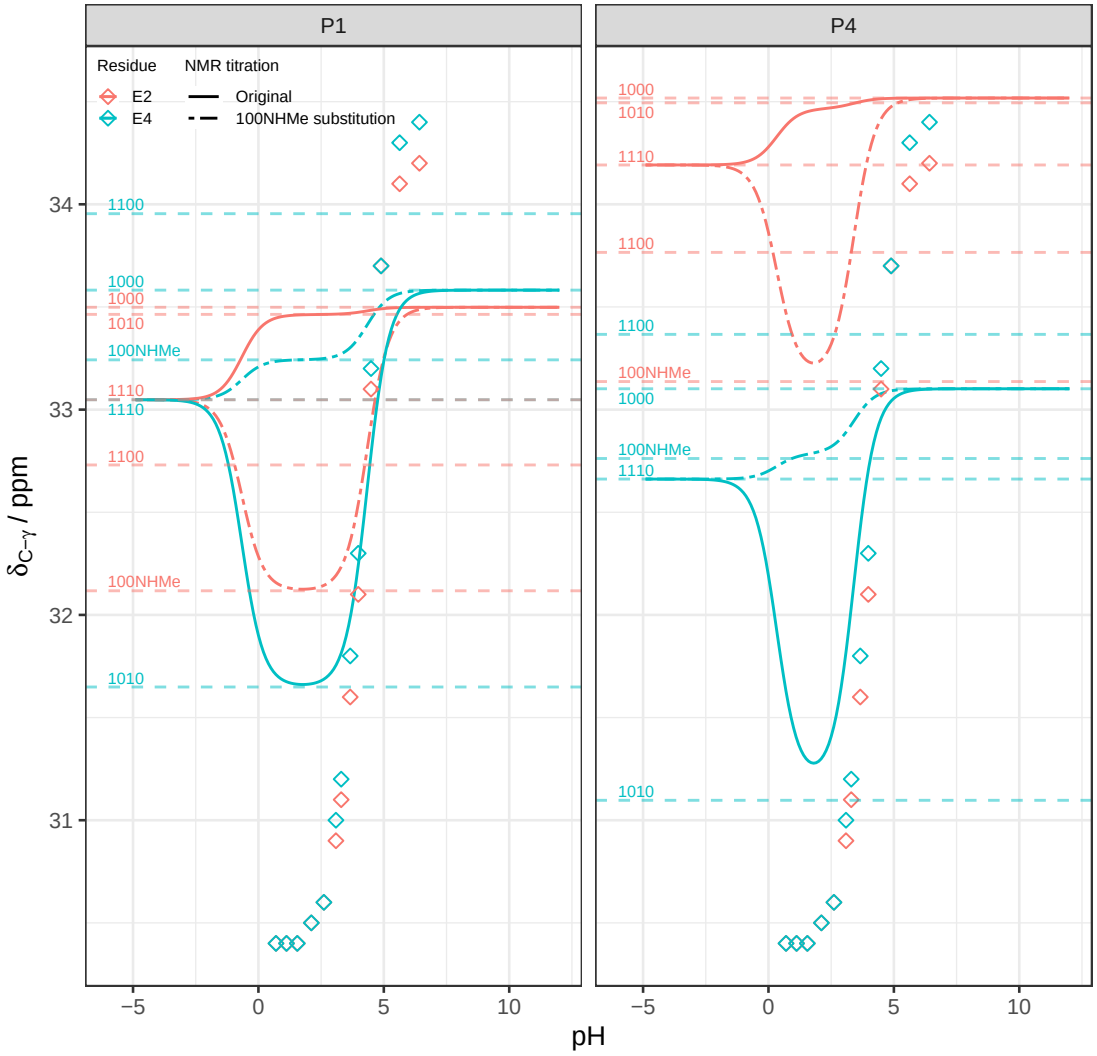

Figure 6.11.: Predicted pH-dependent chemical shift curves (equation 6.11) for the C_γ atoms of the GEAEG pentapeptide from ONIOM-EC-RISM/B calculations with an 4.0 Å explicit solvation shell. The underlying populations were obtained using the *hr*-pK_a correction. Experimental chemical shifts are shown as diamonds. The colours distinguish between the results obtained for either residue E2 (red) or E4 (blue). Dashed horizontal lines show the microstate shifts from tables 6.7 and 6.8

populations.

The resulting curves, shown in figure 6.11, for E2 and both partitions show a clear improvement over the unmodified NMR titration curves. As the populated states 1000 and 1010 showed almost identical shifts, the addition of the C-terminal protonation effect to 1000 causes the newly populated state 100NHMe to give a significantly smaller shift than 1010 and subsequently a clear increase in the weighted shift with increasing pH, as can be observed in the experiment. However, the large experimental range of 3.8 ppm cannot be reproduced.

The same microstate substitution does not improve the prediction quality for the other glutamic acid residue E4. Here the state 1010 already shows a smaller shift than 1000. In this case, adding the C-terminal protonation effect to 1000 leads to a shift for 100NHMe that is larger than the 1010 shift. As a result, the microstate curves show a chemical shift range that is significantly smaller than the experiment and smaller than the range of the unmodified NMR titration curve.

Previously it was observed that extrapolation to the 1110 state by adding the independent protonation effects of the glutamic acid side chains in equation 6.22 gave almost identical shifts to the explicitly calculated state for E2. Larger deviations were observed for E4. As the extrapolation method used is methodologically identical to the approach used to obtain the 100NHMe state, the predictions for E4 may show the same deviations. This could explain the worse predictions obtained for E4 compared to the unsubstituted titration curves. However, this assumption cannot be verified without explicitly simulating the 1001 state.

In addition to these uncertainties, it was shown by comparison with the site-specific pK_a values in section 6.4.3 that the lower of the two macroscopic pK_a values may be too small. It follows that the transition between 1110 and 1010 would occur at higher pH values. Therefore, 1110 would be more abundant at lower pH values. However, the shift for 1110 is significantly larger and would therefore lead to a deterioration in the prediction quality, after all, a state with a significantly smaller shift would have to be populated in this range in order to reproduce the experiment.

Using an approach developed by Ullmann,[130] it was possible to calculate site-specific titration curves for a model diprotic acid. It was shown that for the sigmoidal shape of the NMR titration experiment to be reproduced by the model, the two protonation sites must interact only weakly. In addition, four nearly identical microstate pK_a values are required. However, none of the ONIOM-EC-RISM pK_a models tested here gave such predictions, further suggesting that the microstate energies are not accurately predicted.

Using a reasoning from another publication by Bombarda and Ullmann,[131] it was possible to assume that the lower of the two macro pK_a values is too small and therefore a microstate with C-terminal protonation from the same macrostate as 1110 could also be populated, i.e. 1101 or 1011. These states were artificially constructed using an approximated protonation effect $\Delta_{1110\rightarrow 111\mathrm{NHMe}}\,\delta_{\mathrm{C}_\gamma}$, which for the C-terminal protonation is negative. This would therefore result in a smaller shift and a direct improvement in

the results. However, it is problematic that the construction of the corresponding NMR titration curve requires further manipulation of the underlying population curve in addition to the already speculative approximation used to calculate $\Delta_{1110\rightarrow 111\mathrm{NHMe}}\,\delta_{\mathrm{C}_\gamma}$. The construction of the extrapolated shift of 100NHMe could be motivated on the basis of the results from equation 6.22, but the adjustment of the population curve would additionally require a purely manual manipulation of one of the lower macro-pK_a values, which cannot be reasonably motivated.

It should be clear by now that the study of the C-terminal protonation effect and the resulting substituted NMR titration curve requires a number of approximations. The reliability of the here presented results therefore depends largely on the validity of the underlying approximations. The problem is that many of these approximations cannot be validated without knowing the chemical shift of the reference state, be it 1111 for the 111NHMe approximation or 1001 for 100NHMe. However, due to the initial choice of force field, the same as that used in the reference study, none of these states can be explicitly simulated to check the validity of the introduced approximations.

The obvious and probably most sensible solution to this problem is to replace the force field with one that allows the simulation and explicit calculation of the chemical shifts of all the missing states with a protonated C-terminus. This would, however, require the repetition of all simulations and single point calculations with this new force field, which would result in an enormous computational effort, which cannot be reasonably be done within this work. These states will need to be explored in subsequent studies.

In addition, further efforts are needed to validate the MD-based sampling approach. Although it has been shown that the predicted pK_a values converge with respect to the number of samples, it cannot be excluded that the underlying ensemble is not representative, for example because the pentapeptide explores only a small configurational space around a local minima during the MD simulation. In follow-up studies, special care must be taken to exclude these effects, for example by repeating the simulation used here with different initial geometries.

Although the data suggest an influence of the degree of protonation of the C-terminus on the chemical shifts of the C_γ atoms, it cannot be ruled out that some of the deviations from the experiment are due to the underlying model. This raises the question of how the non-availability of the *low* level shielding due to the implementation and the resulting approximation of the shielding by the *hm* calculation alone affects the prediction quality.

For further investigation of the model quality, future studies should therefore not only rely on experimental results, but also on direct comparison with the *hr*-reference and the size extrapolation limit PFL, as was previously the case for the validation of the method using the SAMPL6 data set. Although in both cases the calculation of these reference values is significantly more costly than the ONIOM-EC-RISM approximation due to the size of the pentapeptide system and the associated *hr* electron structure calculation, the cost can certainly be reduced by using a well-averaged structure or a smaller subset of snapshots.

If these investigations show that the neglect of *low* level chemical shieldings through the use of the EMPIRE code is a significant contributor to the method error, there are two ways to investigate this error further. The first is to re-implement the ONIOM-EC-RISM method, with the *low* level calculations carried out by another suitable quantum-mechanical program which, unlike EMPIRE, allows the calculation of shieldings.

It can be assumed that this will involve a considerable amount of programming and time. To avoid this, the ONIOM-EC-RISM/X approximation can be used. With this second option for a more detailed analysis of the method error, it is possible to take advantage of the fact that the individual sub-calculations are completely independent of each other and that the *low* level calculations can be performed with an already implemented QM code that allows the shieldings to be calculated. This should provide a faster and more suitable method of investigating the error, which should be given priority over the first method, i.e. re-implementing the ONIOM-EC-RISM method by replacing the EMPIRE code.

In summary, the method presented here makes it possible for the first time to model NMR titration experiments on the basis of EC-RISM calculations. However, it should be pointed out that this method is not bound to this solvation model. Any model that is able to calculate both the required macroscopic pK_a values and the microscopic populations from e.g. the free energies of the tautomers, as was done here, and at the same time is able to predict chemical shifts, should be methodologically compatible. It should therefore also be possible to use, for example, PCM or its ONIOM approximation as a basis for predicting titration curves.

7. Summary and outlook

In this work, a novel multiscale solvation model, ONIOM-EC-RISM, was presented. It combines two multiscale approximations: First, a statistical-mechanical approximation of the solvent background based on the 3D-RISM solvation model, and second, a description of the solute by an ONIOM multiscale approximation, with the aim of reducing the overall computational cost while maintaining the accuracy of previous EC-RISM solvation models.

By comparison with a methodologically similar approach based on the polarisable continuum model, a total of four ONIOM-EC-RISM schemes were developed, named ONIOM-EC-RISM/A, /B, /C and /X. In addition to their theoretical outline, their implementation was discussed in detail.

The first scheme, /A, could be derived by applying the ONIOM approximation to both the solute electronic energy and the solute electrostatic potential, resulting in a model that requires the costly *high* level method to be evaluated in each EC-RISM iteration. However, the overall cost in this model can be reduced by restricting the *high* level calculation to a smaller part of the system, i.e. the ONIOM *model* system.

In order to reduce the computational requirements of the /A scheme, the /B scheme approximates the electrostatic potential of the solvents by the more cost effective *low* level calculation. As a result, the convergence of the solvent background can be decoupled from the *model* system calculations required for the ONIOM approximation of the solute electronic energy, thus reducing the total number of expensive *high* level calculations to one.

The /C scheme introduces a further approximation by neglecting the polarising effect of the solvent on the ONIOM *model* system. However, this model offers little or no computational advantage over the /B model. A technical implementation has therefore been omitted.

The last scheme, /X, is conceptually the simplest as it approximates the solute free energy by three independent EC-RISM calculations and performs the ONIOM extrapolation on the results.

After describing their theoretical and technical aspects, the ONIOM-EC-RISM schemes were validated by predicting aqueous $\mathrm{p}K_\mathrm{a}$ values on the SAMPL6 dataset, since both experimental and theoretical reference values obtained from *high* level EC-RISM calculations were available. Previous work has shown that accurate prediction of $\mathrm{p}K_\mathrm{a}$ values requires empirical corrections: One to the free energy of the solute, here called the PMV correction, and one to the resulting free energy difference of the solute and its

conjugated base, which finally yields the pK_a prediction, consequently called the pK_a correction. Given that both corrections have to be parameterised by fitting to experimental data, and the size of these data sets, this would have meant a very large number of manual ONIOM partitions had to be created. Apart from the manual and error-prone work involved, this would have introduced an arbitrary partitioning error. To avoid this, a correction scheme was developed for the ONIOM-EC-RISM model that exploits the size extrapolation limit, here called the partition free limit (PFL), of the ONIOM extrapolation. This removed any partitioning dependence of the resulting parameters.

For the subsequent pK_a predictions, two sets of PMV correction parameters were required. One at the /B@PFL level of theory for calculations with the /B scheme and one at the *lr* level of theory for one of the correction modes of the /X scheme. These PMV corrections were parameterised by predicting free energies of solvation for the aqueous MNSOL data set. The results for the *lr* level of theory, in this case the PM6 Hamiltonian, showed significantly worse predictions than the *hr*, i.e. MP2, reference. In an effort to improve these results, a modification of the linear correction model was devised in which each of the three charge states of the data set, i.e. neutral, anionic and cationic species, was corrected with its own set of parameters. This improved the results, but they were still significantly worse than the *hr* reference. It has been argued that the deviations from the reference are due to the approximate description of the solutes as well as the solvent distribution through the use of the PM6 Hamiltonian.

Increasing the description of the solute to the *hr* level of theory yields the /B@PFL model. Here a significant improvement in prediction quality was observed, but still worse than the *hr* reference. The stepwise increase of the level of approximation from *hr* to /B@PFL to *lr* thus showed that the remaining error of the /B@PFL model is due to the approximation of the *hr* electrostatic potential by the PM6 potential.

For the parameterisation of the pK_a correction with the Klicić data set, predictions were again made at the *lr* and /B@PFL levels of theory. The former yielded results showing large deviations from the prediction quality of the *hr* reference. To improve these results, a similar approach was taken to the parameterisation of the PMV correction. Acids and bases, as defined in the Klicić dataset, were parameterised separately. None of the approaches tested yielded usable results, so all PM6 models were discarded.

Applying the same approaches to the /B@PFL level of theory revealed that large parts of the total prediction error could be attributed to the prediction of the acid subset. Further investigation showed that this was mainly due to the predictions obtained for the substance class "thiols". Therefore, all models were re-parameterised by excluding thiols from the training data set. The resulting models showed a prediction quality that was only slightly worse than the *hr* reference on the full data set, demonstrating that the ONIOM-EC-RISM model is indeed capable of extrapolating to the more expensive *hr*-EC-RISM method, although the inaccuracies for the thiol substance class require further investigation in subsequent studies.

To fully validate ONIOM-EC-RISM, the previously parameterised model had to be

tested on an independent data set. The pK_a data set from the SAMPL6 challenge was chosen for this purpose. However, before the models could be tested, a number of technical challenges had to be overcome. Although the use of the PFL provided a convenient way to parameterise the model, the final model was evaluated using partitioned molecules to assess whether the parameters were transferable to the unpartitioned case. This also meant that partitions had to be defined for each structure in the dataset, resulting in a large amount of manual work due to the large number of macrostates and underlying tautomers and conformers. In order to reduce the overall partitioning effort required, a graph-based algorithm was developed that allows the transfer of a partition from a reference structure to all other tautomers and conformers of the same molecule, thus reducing the number of partitions to be defined to one per molecule.

Given the large number of corrections parameterised, a filtering step was required to identify promising models. As all pK_a models were parameterised for the /B scheme, the initial model selection was performed at this level of theory. Two well performing models were identified, giving prediction qualities equal to or significantly better than the more expensive *hr* reference model on the same data set. Both apply the multipole approximation to the electrostatic potential, as well as the newly parameterised /B@PFL PMV correction parameters, with only one set of parameters describing all charge states. Application of the *hr*-pK_a correction parameters to the resulting free energy difference resulted in deviations from the experimental values with an RMSE of 0.85, which is significantly better than the *hr* reference. Application of the /B@PFL-pK_a correction parameters yielded a higher RSME of 1.13, still as good a prediction quality as the more expensive *hr*-reference model, but surprisingly, the parameters applied are consistently parameterised at the same level of theory, whereas those of the previous models were not. By rigorously investigating the mathematical basis of the final pK_a model, it has been shown that the remarkable results of the non-consistent model can be attributed to the charge correction parameter c_q of the PMV correction. Since it plays a similar role to the offset parameter b of the pK_a correction, it shifted the predicted pK_a values by an amount that fortunately but coincidentally corrected the underprediction observed in the other model.

Furthermore, in order to gain a better understanding of the model, the partitioning effect for the /B scheme was evaluated by comparison with the predictions of the /B@PFL model. It was found that the /B@PFL models gave predictions that were almost identical to the *hr* predictions. Since the /B@PFL model, as well as the empirical correction parameters, are free of any partitioning error, it was shown that the basic approximations of the /B scheme are able to reproduce the accuracy of the more expensive *hr*-EC-RISM pK_a model. It was also shown that the /B@PFL model gave slightly worse results than the corresponding /B models using partitioned molecules. This suggests that the partitioning of molecules has a beneficial effect on the prediction quality of the pK_a models. However, it should be noted that the /B@PFL model still yielded predictions that were almost equal to the *hr*-reference model, thus validating the

basic approximations of this ONIOM-EC-RISM model.

Following the evaluation of the /B scheme, these two pre-selected models, as well as others, were tested for the less approximate /A and /X schemes. For both, excellent predictions were obtained by applying the /B@PFL PMV and pK_a correction. Here the application of both /B@PFL corrections to the /A scheme gave an RSME of 0.78, while the application of the same corrections to the /X scheme gave an RMSE of 0.73. This is an clear improvement to the *hr*-reference model, which gave, previous to this work, the best EC-RISM prediction on the SAMPL6 data set with an RMSE of 1.13. This also clearly demonstrates that the reduction in the level of the ONIOM-EC-RISM approximation leads to improved results.

It was also found that classical corrections in the style of previous works with one set of parameters for all structures, i.e. no separate parameterisation based on charge or substance classes, gave the best results for all schemes /A, /B and /X.

It should be emphasised that the worst predictions obtained from the ONIOM-EC-RISM models using the consistent parameterisation at the /B@PFL level with one set of parameters each, still give the same prediction quality as the *hr* reference. However, especially for the /A and /X schemes, the resulting models clearly outperform the more expensive reference model.

In order to measure the speed-up gained by the multiscale approximations, the runtimes of the ONIOM-EC-RISM schemes were measured and compared with the *hr*- and *lr*-EC-RISM schemes. Here the /B scheme showed the fastest results of all ONIOM-EC-RISM schemes due to the decoupling of the solvent structure calculation from the *high* level of theory. For /A and /X a smaller speed-up was observed, which is particularly pronounced for smaller ONIOM *model* systems. It could be argued that the speed-up is mainly due to the reduction in the number of atoms that need to be treated with the *high* level theory, since in contrast to the /B model the *high* level calculations need to be evaluated in each EC-RISM iteration.

It has therefore been shown that the /B scheme is capable of reproducing the prediction quality of the *hr*-EC-RISM model while drastically reducing the computational effort required. To further test the capabilities of this model, it was applied to a GEAEG pentapeptide system with the aim of predicting acid constants as well as chemical shifts for the glutamic acid side chains, as was done in a publication by Grubmüller and coworkers, which provided computationally predicted pK_a values as well as experimental site-specific pK_a values and chemical shifts obtained from an NMR titration experiment as a reference. In addition to the prediction of pK_a values and chemical shifts, a novel method was developed that allows the direct calculation of pH-dependent NMR titration curves by introducing pH-dependent weights to the microscopically averaged chemical shifts. These weights were calculated directly from microstate populations and macroscopic pK_a values, and thus presented an ideal case to test the combined predictive quality of thermodynamic and spectroscopic quantities.

Unlike the SAMPL6 data set, here the conformers were taken directly from molecular

dynamics simulations using the same force field as in the reference study. While the higher of the two predicted macroscopic pK_a values gave results close to the site-specific reference values, further comparison of the latter with microscopic pK_a values showed that the acidity of the E2 side chain was likely to be overestimated. As a result, the lower of the two macroscopic pK_a values was probably underestimated.

After calculating pH-dependent population curves, the first NMR titration curves could be predicted. Their evaluation showed that the basic ONIOM-EC-RISM model was not able to reproduce the experimental results accurately. On the one hand, both side chains showed shifts significantly larger than the experimental reference, and the range of predicted shifts was too small. In addition, the side chain of E2 showed a decrease in shift with increasing pH, which was the opposite trend observed in the experiment.

To improve the results, explicit solvation shells of 2.5 and 4.0 Å were added, as it has been shown in another publication using EC-RISM that the addition of explicit water molecules improves the prediction of chemical shifts. The addition of increasing numbers of explicit water molecules resulted in a significant reduction in shifts, bringing the predicted shifts into the range observed in the experiment. However, it was evident that the incorrectly predicted trend of the E2 side chain could not be corrected by this approach.

From the data presented it was inferred that the C-terminal carboxy function might influence the protonation of the glutamic acid side chain, and it was decided to investigate this effect further by modelling the previously missing protonated state of the C-terminus. However, as the force field used did not provide adequate parameters, a C-terminal NHMe group was introduced to approximate the fully protonated state of the pentapeptide. The resulting chemical shifts showed that the protonation of the C-terminus mainly affects the E2 side chain. Based on these results, the shifts of another state, which was thought to be more likely to be occupied in the lower pH ranges, were approximated by an additive extrapolation scheme, as it was decided that the explicit simulation of this state should be part of a later study.

As a result, the appropriate populations were unavailable and the weighted NMR titration curve had to be constructed by substituting one state with the newly constructed microstate. The results showed a clear improvement for E2 and a decrease in the predicted shift range for E4, thus worsening the results for the latter residue.

While the method presented here was able to reproduce the general shape of the experimental titration curve, it became apparent that the chemical shifts predicted for the side chain of E2 in particular offer room for improvement and require further investigation. In particular, the investigation of the C-terminal protonation effect required the introduction of approximations, none of which can be reasonably validated without the explicit simulation of the corresponding reference state. However, these reference states are not available, as the chosen force field, the same as in the reference study, does not provide the required parameters for a protonated C-terminus. The only viable option to

solve this problem would have been to change the force field to one that allowed modelling of all the required protonation states, but this would also have meant repeating all previous simulations and calculations for consistency reasons. This problem will need to be addressed in future studies. It should be emphasises that these problems mainly arise due to the force field used in the structure generation, and cannot be attributed to the ONIOM-EC-RISM solvent model.

The present study paves the way for a more in-depth analysis of the pentapeptides and the chemical shift predictions with ONIOM-EC-RISM. One aspect that should be investigated in further studies is the inherent method error due to the neglect of the *low* level shieldings. As these shieldings are not available in EMPIRE, the QM code used for the *low* level calculations, the ONIOM-EC-RISM approximation reduces to an additive scheme. One way to further investigate the resulting method error is by direct comparison with the *hr* reference. Although the size of the solute and the number of structures sampled for each microstate preclude the calculation of *hr* shieldings for the complete set of snapshots, the shieldings can be obtained for a small subset or a well averaged structure.

If these calculations show that there is indeed a large discrepancy between the additive scheme and the reference, two courses of action seem reasonable. Firstly, a different code could be used to allow the calculation of *low* level shieldings. Unfortunately, this would require a reimplementation of the ONIOM-EC-RISM method. Alternatively, the fact that the /X scheme is based on three independent calculations can be exploited, allowing the *low* level calculations to be performed with any code that allows chemical shieldings to be obtained. However, the latter option would significantly increase the computational requirements.

Apart from the further investigation of the pentapeptide, the observed speed-up in comparison to the *hr*-EC-RISM method, especially for the /B scheme, raises the question of how to further reduce the cost of ONIOM-EC-RISM, as this is a prerequisite for the investigation of larger systems such as proteins, if a similar sampling approach as used for the pentapeptide is to be applied. As the measurements clearly showed, the *hm* calculations account for the largest part of the total run time for all schemes. A first reasonable approach would therefore be to replace the MP2 level of theory with a more cost efficient approach.

While this is probably a sensible way of reducing the cost for the smaller systems considered here, for larger systems the computational cost of the *lr* calculations will increase dramatically. It is reasonable to assume that as the size of the system increases, the *model* system should remain similar in size. Hence, there should be a point where the total cost is dominated by the cost of the *lr* and 3D-RISM calculations, as the solvent lattice size also scales with the solute size.

Therefore, if the ONIOM-EC-RISM approach is to be applied to larger systems, e.g. proteins, it is desirable to further reduce the cost of the *lr* and 3D-RISM calculations. One way to achieve this is to replace the semi-empirical QM method with a

non-polarisable force field. This has two effects, firstly the cost of the *lr* calculation is drastically reduced as only one analytical expression needs to be evaluated as opposed to the iterative numerical procedure required for the SQM calculation. As a second effect, the need to perform EC-RISM iterations to polarise the *low* level calculations disappears. Consequently, for the /X scheme, only the solution for the *hm* sub-calculation needs to be obtained by EC-RISM iterations, which greatly reduces the computational cost.

This effect is even more beneficial for the /B scheme. As the calculation of the solvent structure is decoupled from the calculation of the *model* system, the former can be calculated directly using the unpolarisable *lr*-ESP. The energy of the *model* system can now be evaluated directly under the influence of this already converged solvent distribution, thus reducing the number of EC-RISM iterations to just one and potentially offering a way to investigate large system sizes with ONIOM-EC-RISM.

Bibliography

[1] Henry Eyring, John Walter, George Elbert Kimball, *Quantum Chemistry*, John Wiley and Sons Inc., New York, London, **1944**.

[2] Weinan E, *Principles of multiscale modeling*, Cambridge University Press, Cambridge ; New York, **2011**.

[3] T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B.* **2008**, *112*, 4337–4343.

[4] M. Svensson, S. Humbel, K. Morokuma, *J. Chem. Phys.* **1996**, *105*, 3654–3661.

[5] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, K. Morokuma, *Chem. Rev.* **2015**, *115*, 5678–5796.

[6] D. Beglov, B. Roux, *J. Phys. Chem. B* **1997**, *101*, 7821–7826.

[7] A. Kovalenko, S. Ten-no, F. Hirata, *J. Comput. Chem.* **1999**, *20*, 928–936.

[8] Q. Du, D. Beglov, B. Roux, *J. Phys. Chem. B* **2000**, *104*, 796–805.

[9] S. Miertuš, E. Scrocco, J. Tomasi, *Chem. Phys.* **1981**, *55*, 117–129.

[10] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105*, 2999–3094.

[11] T. Vreven, B. Mennucci, C. O. Da Silva, K. Morokuma, J. Tomasi, *J. Chem. Phys.* **2001**, *115*, 62–72.

[12] S. J. Mo, T. Vreven, B. Mennucci, K. Morokuma, J. Tomasi, *Theor. Chem. Acc.* **2004**, *111*, 154–161.

[13] A. Warshel, M. Karplus, *J. Am. Chem. Soc.* **1972**, *94*, 5612–5625.

[14] M. Levitt, A. Warshel, *Nature* **1975**, *253*, 694–698.

[15] M. Levitt, *J. Mol. Biol.* **1976**, *104*, 59–107.

[16] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227–249.

[17] J. Heil, PhD thesis, Technische Universität Dortmund, Dortmund, **2016**.

[18] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, *J. Chem. Phys.* **2020**, *152*, 224108.

[19] F. Neese, *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

[20] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, P. J. Taylor, *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259–279.

[21] S. M. Kast, J. Heil, S. Güssregen, K. F. Schmidt, *J. Comput.-Aided Mol. Des.* **2010**, *24*, 343–353.

[22] C. C. Bannan, K. H. Burley, M. Chiu, M. R. Shirts, M. K. Gilson, D. L. Mobley, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 927.

[23] N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, S. M. Kast, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 1035–1044.

[24] N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1151–1163.

[25] M. Işık, A. S. Rustenburg, A. Rizzi, M. R. Gunner, D. L. Mobley, J. D. Chodera, *J. Comput.-Aided Mol. Des.* **2021**, *35*, 131–166.

[26] N. Tielker, PhD thesis, Technische Universität Dortmund, Dortmund, **2021**.

[27] R. Frach, S. M. Kast, *J. Phys. Chem. A* **2014**, *118*, 11620–11628.

[28] R. Frach, P. Kibies, S. Böttcher, T. Pongratz, S. Strohfeldt, S. Kurrmann, J. Koehler, M. Hofmann, W. Kremer, H. R. Kalbitzer, O. Reiser, D. Horinek, S. M. Kast, *Angew. Chem. Int. Ed.* **2016**, *55*, 8757–8760.

[29] R. Frach, P. Kibies, S. Böttcher, T. Pongratz, S. Strohfeldt, S. Kurrmann, J. Koehler, M. Hofmann, W. Kremer, H. R. Kalbitzer, O. Reiser, D. Horinek, S. M. Kast, *Angew. Chem. Int. Ed.* **2016**, *55*, 11713–11713.

[30] C. E. Munte, M. Karl, W. Kauter, L. Eberlein, T.-V. Pham, M. B. Erlach, S. M. Kast, W. Kremer, H. R. Kalbitzer, *Biophys. Chem.* **2019**, *254*, 106261.

[31] C. E. Munte, M. Karl, W. Kauter, L. Eberlein, T.-V. Pham, M. Beck Erlach, S. M. Kast, W. Kremer, H. R. Kalbitzer, *Biophys. Chem.* **2020**, *265*, 106408.

[32] 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.* **2020**, *257*, 106258.

[33] S. Maste, B. Sharma, T. Pongratz, B. Grabe, W. Hiller, M. Beck Erlach, W. Kremer, H. Robert Kalbitzer, D. Marx, S. M. Kast, *Phys. Chem. Chem. Phys.* **2024**, *26*, 6386–6395.

[34] R. Frach, PhD thesis, Technische Universität Dortmund, Dortmund, **2015**.

[35] T. Pongratz, PhD thesis, Technische Universität Dortmund, Dortmund, **2022**.

[36] H. M. Senn, W. Thiel, *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.

[37] H. M. Senn, W. Thiel, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[38] C. M. Clemente, L. Capece, M. A. Martí, *J. Chem. Inf. Model.* **2023**, *63*, 2609–2627.

[39] K.-S. Csizi, M. Reiher, *WIREs Comput. Mol. Sci.* **2023**, *13*, e1656.

[40] D. Bakowies, W. Thiel, *J. Phys. Chem.* **1996**, *100*, 10580–10594.

[41] H. M. Senn, W. Thiel in *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*, (Ed.: M. Reiher), Topics in Current Chemistry, Springer, Berlin, Heidelberg, **2007**, pp. 173–290.

[42] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian 16 Rev. C.01, Wallingford, CT, **2016**.

[43] M. Bondanza, M. Nottoli, L. Cupellini, F. Lipparini, B. Mennucci, *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433–14448.

[44] C. N. Rowley, B. Roux, *J. Chem. Theory Comput.* **2012**, *8*, 3526–3535.

[45] E. Boulanger, W. Thiel, *J. Chem. Theory Comput.* **2012**, *8*, 4527–4538.

[46] J. A. Lemkul, J. Huang, B. Roux, A. D. J. MacKerell, *Chem. Rev.* **2016**, *116*, 4983–5013.

[47] S. K. Sahoo, N. N. Nair, *Front. Chem.* **2018**, *6*.

[48] J. Dziedzic, Y. Mao, Y. Shao, J. Ponder, T. Head-Gordon, M. Head-Gordon, C.-K. Skylaris, *J. Chem. Phys.* **2016**, *145*, 124106.

[49] D. Loco, É. Polack, S. Caprasecca, L. Lagardère, F. Lipparini, J.-P. Piquemal, B. Mennucci, *J. Chem. Theory Comput.* **2016**, *12*, 3654–3661.

[50] C. Curutchet, A. Muñoz-Losa, S. Monti, J. Kongsted, G. D. Scholes, B. Mennucci, *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848.

[51] J. Gao, *J. Comput. Chem.* **1997**, *18*, 1061–1071.

[52] M. A. Thompson, G. K. Schenter, *J. Phys. Chem.* **1995**, *99*, 6374–6386.

[53] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, L. V. Slipchenko, *Chem. Rev.* **2012**, *112*, 632–672.

[54] J. Gao, D. G. Truhlar, Y. Wang, M. J. M. Mazack, P. Löffler, M. R. Provorse, P. Rehak, *Acc. Chem. Res.* **2014**, *47*, 2837–2845.

[55] A. O. Dohn, *Int. J. Quantum Chem.* **2020**, *120*, e26343.

[56] S. Dapprich, I. Komáromi, K. Byun, K. Morokuma, M. J. Frisch, *J. Mol. Struct. THEOCHEM* **1999**, *461-462*, 1–21.

[57] U. Eichler, C. M. Klmel, J. Sauer, *J. Comput. Chem.* **1997**, *18*, 463–477.

[58] T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. Montgomery, K. Morokuma, M. J. Frisch, *J. Chem. Theory Comput.* **2006**, *2*, 815–826.

[59] D. Das, K. P. Eurenius, E. M. Billings, P. Sherwood, D. C. Chatfield, M. Hodošček, B. R. Brooks, *J. Chem. Phys.* **2002**, *117*, 10534–10547.

[60] X.-P. Wu, L. Gagliardi, D. G. Truhlar, *Molecules* **2018**, *23*, 1309.

[61] Y. Zhang, T.-S. Lee, W. Yang, *J. Chem. Phys.* **1999**, *110*, 46–54.

[62] Y. Zhang, *J. Chem. Phys.* **2004**, *122*, 024114.

[63] D. M. Philipp, R. A. Friesner, *J. Comput. Chem.* **1999**, *20*, 1468–1494.

[64] V. Kairys, J. H. Jensen, *J. Phys. Chem. A* **2000**, *104*, 6656–6665.

[65] V. Gogonea, L. M. Westerhoff, K. M. Merz, *J. Chem. Phys.* **2000**, *113*, 5604–5613.

[66] F. Maseras, K. Morokuma, *J. Comput. Chem.* **1995**, *16*, 1170–1179.

[67] S. Humbel, S. Sieber, K. Morokuma, *J. Chem. Phys.* **1996**, *105*, 1959–1967.

[68] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, K. Morokuma, *J. Phys. Chem.* **1996**, *100*, 19357–19363.

[69] M. A. Collins, R. P. A. Bettens, *Chem. Rev.* **2015**, *115*, 5607–5642.

[70] S. Caprasecca, S. Jurinovich, L. Viani, C. Curutchet, B. Mennucci, *J. Chem. Theory Comput.* **2014**, *10*, 1588–1598.

[71] M. F. S. J. Menger, S. Caprasecca, B. Mennucci, *J. Chem. Theory Comput.* **2017**, *13*, 3778–3786.

[72] J. Tomasi, B. Mennucci, E. Cancès, *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211–226.

[73] B. Mennucci, *WIREs Comput. Mol. Sci.* **2012**, *2*, 386–404.

[74] G. Scalmani, M. J. Frisch, *J. Chem. Phys.* **2010**, *132*, 114110.

[75] D. M. York, M. Karplus, *J. Phys. Chem. A* **1999**, *103*, 11060–11079.

[76] M. Cossi, V. Barone, R. Cammi, J. Tomasi, *Chem. Phys. Lett.* **1996**, *255*, 327–335.

[77] J. Hansen, I. McDonald, *Theory of simple liquids*, 3rd ed., (Eds.: J.-P. Hansen, I. R. McDonald), Academic Press, Burlington, **2006**.

[78] F. Hirata in *Molecular Theory of Solvation*, *Vol. 24*, (Ed.: F. Hirata), Kluwer Academic Publishers, Dordrecht, **2004**, pp. 1–60.

[79] F. Hirata, *Exploring life phenomena with statistical mechanics of molecular liquids*, CRC Press, Boca Raton, **2020**.

[80] J. S. Rowlinson, *Rep. Prog. Phys.* **1965**, *28*, 169–199.

[81] M. Klein, M. S. Green, *J. Chem. Phys.* **1963**, *39*, 1367–1387.

[82] P. Attard, G. N. Patey, *J. Chem. Phys.* **1990**, *92*, 4970–4982.

[83] G. D. J. Phillies, *J. Stat. Phys.* **1982**, *28*, 673–683.

[84] H. C. Andersen, D. Chandler, *J. Chem. Phys.* **1972**, *57*, 1918–1929.

[85] D. Chandler, H. C. Andersen, *J. Chem. Phys.* **1972**, *57*, 1930–1937.

[86] A. Kovalenko, F. Hirata, *J. Chem. Phys.* **1999**, *110*, 10095–10112.

[87] F. Hirata, P. J. Rossky, *Chem. Phys. Lett.* **1981**, *83*, 329–334.

[88] F. Hirata, B. M. Pettitt, P. J. Rossky, *J. Chem. Phys.* **1982**, *77*, 509–520.

[89] F. Hirata, P. J. Rossky, B. M. Pettitt, *J. Chem. Phys.* **1983**, *78*, 4133–4144.

[90] J. S. Perkyns, B. Montgomery Pettitt, *Chem. Phys. Lett.* **1992**, *190*, 626–630.

[91] J. Perkyns, B. M. Pettitt, *J. Chem. Phys.* **1992**, *97*, 7656–7666.

[92] J. Perkyns, B. Montgomery Pettitt, *J. Chem. Phys.* **1994**, *100*, 8556.

[93] A. Kovalenko, F. Hirata, *J. Phys. Chem. B* **1999**, *103*, 7942–7957.

[94] S. M. Kast, *Phys. Rev. E* **2003**, *67*, 041203.

[95] S. M. Kast, *Phys. Rev. E* **2006**, *73*, 012201.

[96] S. M. Kast, T. Kloss, *J. Chem. Phys.* **2008**, *129*, 236101.

[97] J. Heil, S. M. Kast, *J. Chem. Phys.* **2015**, *142*, 114107.

[98] Y. Harano, T. Imai, A. Kovalenko, M. Kinoshita, F. Hirata, *J. Chem. Phys.* **2001**, *114*, 9506–9511.

[99] T. Imai, Y. Harano, A. Kovalenko, F. Hirata, *Biopolymers* **2001**, *59*, 512–519.

[100] E. L. Ratkova, D. S. Palmer, M. V. Fedorov, *Chem. Rev.* **2015**, *115*, 6312–6356.

[101] M. Reimann, M. Kaupp, *J. Phys. Chem. A* **2020**, *124*, 7439–7452.

[102] S. Ten-no, F. Hirata, S. Kato, *Chem. Phys. Lett.* **1993**, *214*, 391–396.

[103] H. Sato, F. Hirata, S. Kato, *J. Chem. Phys.* **1996**, *105*, 1546–1551.

[104] H. Sato, A. Kovalenko, F. Hirata, *J. Chem. Phys.* **2000**, *112*, 9463–9468.

[105] Q. Du, D. Wei, *J. Phys. Chem. B* **2003**, *107*, 13463–13470.

[106] Q. Du, D. Beglov, D. Wei, B. Roux, *J. Phys. Chem. B* **2007**, *111*, 13658–13658.

[107] K. Imamura, D. Yokogawa, H. Sato, *J. Chem. Phys.* **2024**, *160*, 050901.

[108] M. Hennemann, J. Margraf, B. Meyer, T. Clark, EMPIRE20, Obermichelbach, **2020**.

[109] D. S. Palmer, A. I. Frolov, E. L. Ratkova, M. V. Fedorov, *Mol. Pharm.* **2011**, *8*, 1423–1429.

[110] V. Sergiievskyi, G. Jeanmairet, M. Levesque, D. Borgis, *J. Chem. Phys.* **2015**, *143*, 184116.

[111] T. Luchko, N. Blinov, G. C. Limon, K. P. Joyce, A. Kovalenko, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 1115–1127.

[112] V. P. Sergiievskyi, G. Jeanmairet, M. Levesque, D. Borgis, *J. Phys. Chem. Lett.* **2014**, *5*, 1935–1942.

[113] M. Misin, M. V. Fedorov, D. S. Palmer, *J. Phys. Chem. B* **2016**, *120*, 975–983.

[114] M. Hennemann, T. Clark, *J. Mol. Model.* **2014**, *20*, 2331.

[115] C. R. Wick, M. Hennemann, J. J. P. Stewart, T. Clark, *J. Mol. Model.* **2014**, *20*, 2159.

[116] L. Eisel, Electronic appendix to thesis: "Multiscale approximations of integral equation-based solvation models", TUDOdata, https://doi.org/10.17877/TUDODATA-2025-MBQCBF48, **2025**.

[117] A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer, D. G. Truhlar, Minnesota Solvation Database – version 2012, Minneapolis, **2012**.

[118] J. J. Klicić, R. A. Friesner, S.-Y. Liu, W. C. Guida, *J. Phys. Chem. A* **2002**, *106*, 1327–1335.

[119] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[120] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem.* **1987**, *91*, 6269–6271.

[121] J. C. Kromann, C. Steinmann, J. H. Jensen, *J. Chem. Phys.* **2018**, *149*, 104102.

[122] N. Tielker, L. Eberlein, C. Chodun, S. Güssregen, S. M. Kast, *J. Mol. Model.* **2019**, *25*, 139.

[123] B. D. McKay, A. Piperno, *J. Symb. Comput.* **2014**, *60*, 94–112.

[124] P. Dobrev, S. P. B. Vemulapalli, N. Nath, C. Griesinger, H. Grubmüller, *J. Chem. Theory Comput.* **2020**, *16*, 2561–2569.

[125] P. Dobrev, S. P. B. Vemulapalli, N. Nath, C. Griesinger, G. Groenhof, H. Grubmüller, *J. Chem. Theory Comput.* **2020**, *16*, 4753–4753.

[126] S. Donnini, F. Tegeler, G. Groenhof, H. Grubmüller, *J. Chem. Theory Comput.* **2011**, *7*, 1962–1978.

[127] S. Donnini, F. Tegeler, G. Groenhof, H. Grubmüller, *J. Chem. Theory Comput.* **2013**, *9*, 3261–3261.

[128] S. Donnini, R. T. Ullmann, G. Groenhof, H. Grubmüller, *J. Chem. Theory Comput.* **2016**, *12*, 1040–1051.

[129] P. Dobrev, S. Donnini, G. Groenhof, H. Grubmüller, *J. Chem. Theory Comput.* **2017**, *13*, 147–160.

[130] G. M. Ullmann, *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

[131] E. Bombarda, G. M. Ullmann, *J. Phys. Chem. B* **2010**, *114*, 1994–2003.

[132] N. C. Handy, A. J. Cohen, *Mol. Phys.* **2001**, *99*, 403–412.

[133] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **2006**, *65*, 712–725.

[134] H. J. C. Berendsen, D. van der Spoel, R. van Drunen, *Comput. Phys. Commun.* **1995**, *91*, 43–56.

[135] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, *J. Comput. Chem.* **2005**, *26*, 1701–1718.

[136] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1-2*, 19–25.

[137] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[138] S. Miyamoto, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 952–962.

[139] S. Nosé, *Mol. Phys.* **1984**, *52*, 255–268.

[140] W. G. Hoover, *Phys. Rev. A* **1985**, *31*, 1695–1697.

[141] M. Parrinello, A. Rahman, *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[142] S. Nosé, M. L. Klein, *Mol. Phys.* **1983**.

[143] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[144] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[145] N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des.* **2020**, *34*, 453–461.

[146] G. M. Ullmann, E.-W. Knapp, *Eur. Biophys. J.* **1999**, *28*, 533–551.

[147] A. Onufriev, D. A. Case, G. M. Ullmann, *Biochemistry* **2001**, *40*, 3413–3419.

[148] C. Tanford, R. Roxby, *Biochemistry* **1972**, *11*, 2192–2198.

[149] R. Fraczkiewicz, M. Waldman, *J. Chem. Inf. Model.* **2023**, *63*, 3198–3208.

[150] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[151] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, **2016**.

List of Figures

List of Tables

Part IV.

Appendix

Additional data

Additional data referenced in the main text are given below. Raw data such as atomic coordinates and the parsed output of single point calculations can be found in the electronic supplementary material.[116] It can be accessed under `https://doi.org/10.17877/TUDODATA-2025-MBQCBF48`. For a full description of its content, see this link.

1. Prediction of solvation free energies for the MNSOL data set

1.1. Comparison of excess chemical potentials at different convergence thresholds

Table 1.: Statistical quantities calculated for the difference $\Delta\mu_{\mathrm{ex}} = \mu_{\mathrm{ex}}^{\mathrm{lr}} - \mu_{\mathrm{ex}}^{/\mathrm{B@PFL}}$ for the MNSOL water subset from *lr*-EC-RISM and ONIOM-EC-RISM/B@PFL calculations, both performed with PM6 as the low-level theory. All values are given in kcal mol^{-1}. See section 5.1.2 for further context.

Opt.	Potential	RMSE	MAE	MSE
B3LYP-PCM	Exact (Mult.)	0.00859	0.00520	0.00519
B3LYP-PCM	Exact (NDDO)	0.00659	0.00492	0.00453
B3LYP-PCM	Point charge	0.00593	0.00389	0.00376
PM6-PCM	Exact (Mult.)	0.00825	0.00522	0.00515
PM6-PCM	Exact (NDDO)	0.00710	0.00495	0.00452
PM6-PCM	Point charge	0.00568	0.00388	0.00370

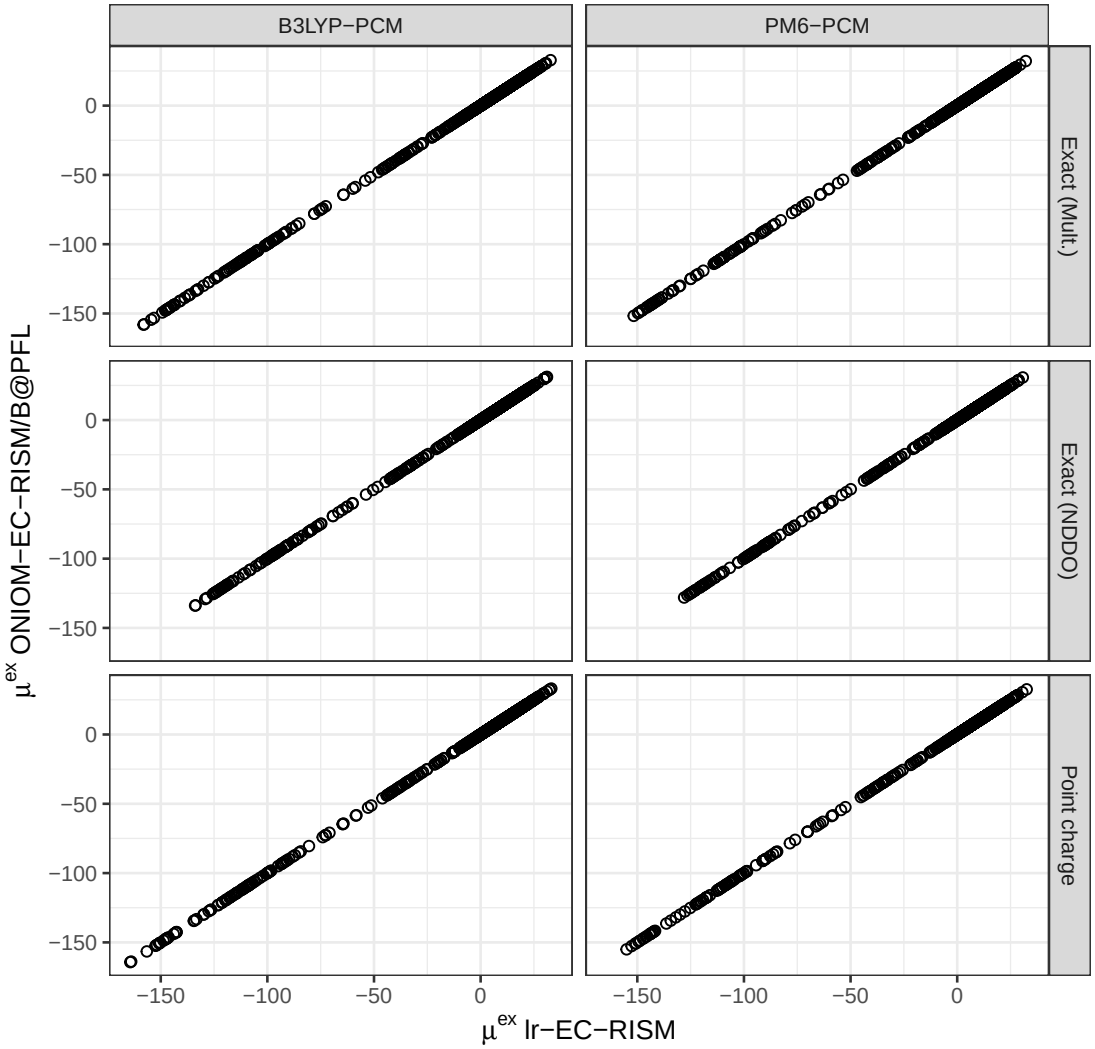

Figure 1.: Comparison of the excess chemical potential for the MNSOL water subset obtained from *lr*-EC-RISM and ONIOM-EC-RISM/B@PFL calculations, both performed with PM6 as the low-level theory. The columns represent the level of theory used to optimise the geometry of the structures, while the rows represent the potential supplied to the 3D-RISM solver. See section 5.1.2 for more context on these calculations.

2. Prediction of acidity constants for the Klicić data set

2.1. Comparison of excess chemical potentials at different convergence thresholds

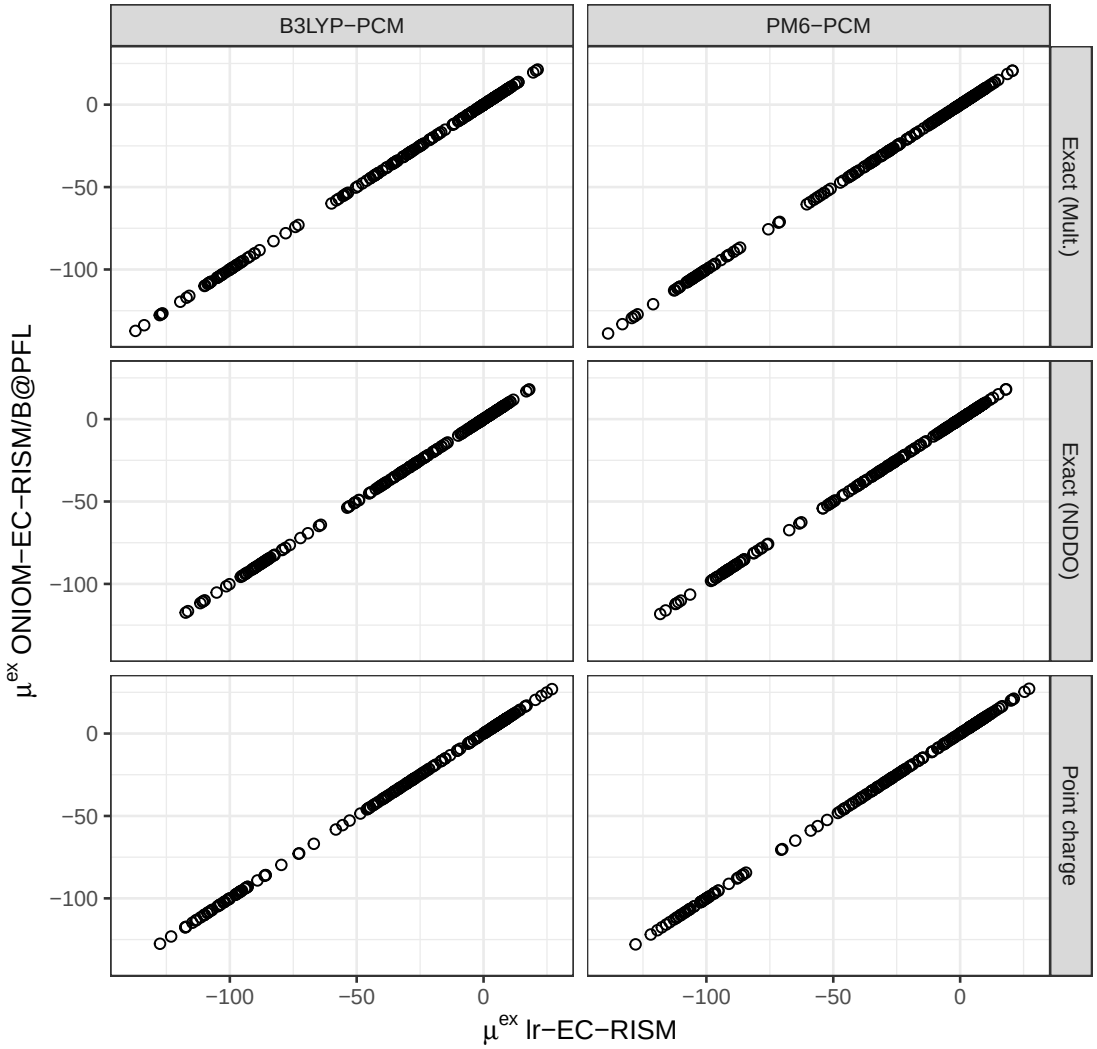

Figure 2.: Comparison of the excess chemical potential for the Klicić data set obtained from *lr*-EC-RISM and ONIOM-EC-RISM/B@PFL calculations, both performed with PM6 as the low-level theory. The columns represent the level of theory used to optimise the geometry of the structures, while the rows represent the potential supplied to the 3D-RISM solver. See section 5.2.2 for more context on these calculations.

Table 2.: Statistical quantities calculated for the difference $\Delta\mu_{\mathrm{ex}} = \mu_{\mathrm{ex}}^{\mathrm{lr}} - \mu_{\mathrm{ex}}^{/\mathrm{B@PFL}}$ for the Klicić data set from *lr*-EC-RISM and ONIOM-EC-RISM/B@PFL calculations, both performed with PM6 as the low-level theory. All values are given in $\mathrm{kcal\,mol^{-1}}$. See section 5.2.2 for further context.

opt	potential	RMSE	MAE	MSE
B3LYP-PCM	Exact (Mult.)	0.01193	0.00621	0.00620
B3LYP-PCM	Exact (NDDO)	0.00692	0.00524	0.00475
B3LYP-PCM	Point charge	0.00656	0.00417	0.00386
PM6-PCM	Exact (Mult.)	0.01140	0.00612	0.00610
PM6-PCM	Exact (NDDO)	0.00667	0.00511	0.00510
PM6-PCM	Point charge	0.00603	0.00396	0.00333

2.2. Statistical values per molecular class

Table 3.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the multipole based ESP on PM6 PCM reoptimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (Mult.)	All	All	acids	0.99	0.80	0.09
Exact (Mult.)	All	All	amines	0.92	0.76	-0.18
Exact (Mult.)	All	All	anilines	1.58	1.45	1.45
Exact (Mult.)	All	All	heterocycles	1.28	1.12	0.40
Exact (Mult.)	All	All	indoles	2.89	2.72	2.72
Exact (Mult.)	All	All	phenoles	1.00	0.85	-0.00
Exact (Mult.)	All	All	pyrroles	3.23	3.14	3.14
Exact (Mult.)	All	All	thiols	7.38	7.26	-7.26
Exact (Mult.)	All	Individual	acids	1.31	1.01	0.96
Exact (Mult.)	All	Individual	amines	0.96	0.75	0.09
Exact (Mult.)	All	Individual	anilines	0.78	0.67	0.46
Exact (Mult.)	All	Individual	heterocycles	1.33	0.94	-0.56
Exact (Mult.)	All	Individual	indoles	0.80	0.72	0.18
Exact (Mult.)	All	Individual	phenoles	1.43	1.27	-0.98
Exact (Mult.)	All	Individual	pyrroles	1.09	0.93	0.93
Exact (Mult.)	All	Individual	thiols	6.04	5.91	-5.91
Exact (Mult.)	Individual	All	acids	1.45	1.16	0.40
Exact (Mult.)	Individual	All	amines	1.06	0.91	-0.25
Exact (Mult.)	Individual	All	anilines	1.18	1.00	1.00
Exact (Mult.)	Individual	All	heterocycles	1.23	1.04	-0.21
Exact (Mult.)	Individual	All	indoles	2.28	2.11	2.11
Exact (Mult.)	Individual	All	phenoles	1.08	0.77	0.33
Exact (Mult.)	Individual	All	pyrroles	2.24	2.16	2.16
Exact (Mult.)	Individual	All	thiols	6.30	6.19	-6.19
Exact (Mult.)	Individual	Individual	acids	1.46	1.16	0.90
Exact (Mult.)	Individual	Individual	amines	1.09	0.89	0.09
Exact (Mult.)	Individual	Individual	anilines	0.84	0.69	0.52
Exact (Mult.)	Individual	Individual	heterocycles	1.44	1.03	-0.69
Exact (Mult.)	Individual	Individual	indoles	1.01	0.94	0.68
Exact (Mult.)	Individual	Individual	phenoles	1.44	1.21	-1.01
Exact (Mult.)	Individual	Individual	pyrroles	0.98	0.86	0.86
Exact (Mult.)	Individual	Individual	thiols	5.52	5.44	-5.44

Table 4.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the NDDO based ESP on PM6 PCM reoptimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (NDDO)	All	All	acids	2.45	2.34	2.34
Exact (NDDO)	All	All	amines	2.62	2.46	-2.46
Exact (NDDO)	All	All	anilines	1.37	1.18	-1.18
Exact (NDDO)	All	All	heterocycles	1.93	1.63	-1.63
Exact (NDDO)	All	All	indoles	0.96	0.88	0.27
Exact (NDDO)	All	All	phenoles	1.79	1.66	1.66
Exact (NDDO)	All	All	pyrroles	1.21	0.92	0.92
Exact (NDDO)	All	All	thiols	3.30	3.19	-3.19
Exact (NDDO)	All	Individual	acids	0.97	0.79	0.67
Exact (NDDO)	All	Individual	amines	0.94	0.76	0.07
Exact (NDDO)	All	Individual	anilines	0.64	0.56	-0.05
Exact (NDDO)	All	Individual	heterocycles	1.00	0.75	-0.28
Exact (NDDO)	All	Individual	indoles	0.75	0.54	-0.16
Exact (NDDO)	All	Individual	phenoles	0.69	0.59	-0.22
Exact (NDDO)	All	Individual	pyrroles	1.09	0.91	0.91
Exact (NDDO)	All	Individual	thiols	4.95	4.88	-4.88
Exact (NDDO)	Individual	All	acids	0.90	0.74	0.04
Exact (NDDO)	Individual	All	amines	0.97	0.74	0.15
Exact (NDDO)	Individual	All	anilines	0.92	0.76	0.64
Exact (NDDO)	Individual	All	heterocycles	0.99	0.89	0.17
Exact (NDDO)	Individual	All	indoles	1.61	1.41	1.41
Exact (NDDO)	Individual	All	phenoles	0.75	0.51	0.00
Exact (NDDO)	Individual	All	pyrroles	2.04	1.95	1.95
Exact (NDDO)	Individual	All	thiols	5.00	4.92	-4.92
Exact (NDDO)	Individual	Individual	acids	1.00	0.80	0.60
Exact (NDDO)	Individual	Individual	amines	0.98	0.76	0.10
Exact (NDDO)	Individual	Individual	anilines	0.66	0.59	0.01
Exact (NDDO)	Individual	Individual	heterocycles	1.07	0.79	-0.40
Exact (NDDO)	Individual	Individual	indoles	0.72	0.67	0.15
Exact (NDDO)	Individual	Individual	phenoles	0.71	0.53	-0.16
Exact (NDDO)	Individual	Individual	pyrroles	0.98	0.82	0.82
Exact (NDDO)	Individual	Individual	thiols	4.53	4.46	-4.46

Table 5.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the point charge based ESP on PM6 PCM reoptimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Point charge	All	All	acids	1.21	1.00	0.12
Point charge	All	All	amines	0.83	0.74	0.28
Point charge	All	All	anilines	0.91	0.76	0.76
Point charge	All	All	heterocycles	1.21	0.96	0.41
Point charge	All	All	indoles	1.87	1.63	1.63
Point charge	All	All	phenoles	1.66	1.40	-1.11
Point charge	All	All	pyrroles	2.69	2.58	2.58
Point charge	All	All	thiols	6.23	6.07	-6.07
Point charge	All	Individual	acids	1.44	1.09	0.86
Point charge	All	Individual	amines	0.82	0.74	0.10
Point charge	All	Individual	anilines	0.43	0.34	-0.14
Point charge	All	Individual	heterocycles	1.23	0.83	-0.39
Point charge	All	Individual	indoles	0.82	0.72	-0.03
Point charge	All	Individual	phenoles	1.49	1.29	-0.87
Point charge	All	Individual	pyrroles	1.32	1.14	1.14
Point charge	All	Individual	thiols	5.46	5.30	-5.30
Point charge	Individual	All	acids	1.38	1.07	0.24
Point charge	Individual	All	amines	0.83	0.66	0.33
Point charge	Individual	All	anilines	0.71	0.56	0.55
Point charge	Individual	All	heterocycles	1.12	0.84	0.11
Point charge	Individual	All	indoles	1.64	1.40	1.40
Point charge	Individual	All	phenoles	1.63	1.35	-1.07
Point charge	Individual	All	pyrroles	2.16	2.06	2.06
Point charge	Individual	All	thiols	5.62	5.46	-5.46
Point charge	Individual	Individual	acids	1.54	1.16	0.81
Point charge	Individual	Individual	amines	0.79	0.61	0.15
Point charge	Individual	Individual	anilines	0.43	0.37	-0.13
Point charge	Individual	Individual	heterocycles	1.25	0.77	-0.50
Point charge	Individual	Individual	indoles	0.81	0.71	0.23
Point charge	Individual	Individual	phenoles	1.50	1.26	-0.89
Point charge	Individual	Individual	pyrroles	1.16	1.01	1.01
Point charge	Individual	Individual	thiols	5.06	4.90	-4.90

Table 6.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the multipole based ESP on B3LYP-PCM-optimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (Mult.)	All	All	acids	0.88	0.67	0.38
Exact (Mult.)	All	All	amines	0.91	0.69	-0.01
Exact (Mult.)	All	All	anilines	0.99	0.87	0.87
Exact (Mult.)	All	All	heterocycles	1.21	1.07	0.46
Exact (Mult.)	All	All	indoles	3.02	2.91	2.91
Exact (Mult.)	All	All	phenoles	1.92	1.77	-1.46
Exact (Mult.)	All	All	pyrroles	3.46	3.40	3.40
Exact (Mult.)	All	All	thiols	8.01	7.95	-7.95
Exact (Mult.)	All	Individual	acids	1.44	1.23	1.22
Exact (Mult.)	All	Individual	amines	0.95	0.70	0.16
Exact (Mult.)	All	Individual	anilines	0.65	0.57	-0.36
Exact (Mult.)	All	Individual	heterocycles	1.26	0.88	-0.55
Exact (Mult.)	All	Individual	indoles	0.74	0.71	0.43
Exact (Mult.)	All	Individual	phenoles	2.42	2.14	-2.09
Exact (Mult.)	All	Individual	pyrroles	1.32	1.24	1.24
Exact (Mult.)	All	Individual	thiols	6.29	6.21	-6.21
Exact (Mult.)	Individual	All	acids	1.22	0.98	0.64
Exact (Mult.)	Individual	All	amines	0.95	0.75	0.04
Exact (Mult.)	Individual	All	anilines	0.64	0.53	0.41
Exact (Mult.)	Individual	All	heterocycles	1.14	0.98	-0.08
Exact (Mult.)	Individual	All	indoles	2.37	2.27	2.27
Exact (Mult.)	Individual	All	phenoles	1.73	1.58	-1.20
Exact (Mult.)	Individual	All	pyrroles	2.49	2.44	2.44
Exact (Mult.)	Individual	All	thiols	7.25	7.17	-7.17
Exact (Mult.)	Individual	Individual	acids	1.48	1.22	1.18
Exact (Mult.)	Individual	Individual	amines	0.98	0.75	0.18
Exact (Mult.)	Individual	Individual	anilines	0.65	0.61	-0.30
Exact (Mult.)	Individual	Individual	heterocycles	1.38	1.00	-0.68
Exact (Mult.)	Individual	Individual	indoles	1.03	0.88	0.84
Exact (Mult.)	Individual	Individual	phenoles	2.35	2.07	-2.01
Exact (Mult.)	Individual	Individual	pyrroles	1.21	1.15	1.15
Exact (Mult.)	Individual	Individual	thiols	6.03	5.98	-5.98

Table 7.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the NDDO based ESP on B3LYP-PCM-optimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (NDDO)	All	All	acids	2.75	2.64	2.64
Exact (NDDO)	All	All	amines	2.58	2.42	-2.41
Exact (NDDO)	All	All	anilines	1.89	1.82	-1.82
Exact (NDDO)	All	All	heterocycles	1.92	1.64	-1.64
Exact (NDDO)	All	All	indoles	0.92	0.88	0.47
Exact (NDDO)	All	All	phenoles	0.82	0.65	0.63
Exact (NDDO)	All	All	pyrroles	1.37	1.17	1.17
Exact (NDDO)	All	All	thiols	3.83	3.79	-3.79
Exact (NDDO)	All	Individual	acids	1.21	1.04	0.94
Exact (NDDO)	All	Individual	amines	0.94	0.69	0.12
Exact (NDDO)	All	Individual	anilines	0.98	0.85	-0.85
Exact (NDDO)	All	Individual	heterocycles	1.01	0.78	-0.28
Exact (NDDO)	All	Individual	indoles	0.59	0.57	0.16
Exact (NDDO)	All	Individual	phenoles	1.29	1.17	-1.17
Exact (NDDO)	All	Individual	pyrroles	1.37	1.28	1.28
Exact (NDDO)	All	Individual	thiols	5.50	5.47	-5.47
Exact (NDDO)	Individual	All	acids	0.92	0.74	0.16
Exact (NDDO)	Individual	All	amines	0.96	0.72	0.35
Exact (NDDO)	Individual	All	anilines	0.54	0.50	0.20
Exact (NDDO)	Individual	All	heterocycles	1.06	0.92	0.44
Exact (NDDO)	Individual	All	indoles	2.13	2.03	2.03
Exact (NDDO)	Individual	All	phenoles	1.58	1.51	-1.51
Exact (NDDO)	Individual	All	pyrroles	2.72	2.66	2.66
Exact (NDDO)	Individual	All	thiols	5.97	5.91	-5.91
Exact (NDDO)	Individual	Individual	acids	1.22	1.00	0.89
Exact (NDDO)	Individual	Individual	amines	0.92	0.69	0.15
Exact (NDDO)	Individual	Individual	anilines	0.95	0.82	-0.80
Exact (NDDO)	Individual	Individual	heterocycles	1.08	0.82	-0.38
Exact (NDDO)	Individual	Individual	indoles	0.68	0.66	0.38
Exact (NDDO)	Individual	Individual	phenoles	1.21	1.10	-1.10
Exact (NDDO)	Individual	Individual	pyrroles	1.27	1.19	1.19
Exact (NDDO)	Individual	Individual	thiols	5.21	5.17	-5.17

Table 8.: Statistical quantities RMSE, MSE and MAE for each molecular class from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B@PFL with the point charge based ESP on B3LYP-PCM-optimised geometries from the Klicić dataset.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Point charge	All	All	acids	1.12	0.90	0.47
Point charge	All	All	amines	0.89	0.67	0.37
Point charge	All	All	anilines	0.42	0.39	0.10
Point charge	All	All	heterocycles	1.12	0.92	0.41
Point charge	All	All	indoles	1.96	1.79	1.79
Point charge	All	All	phenoles	2.83	2.54	-2.45
Point charge	All	All	pyrroles	2.89	2.81	2.81
Point charge	All	All	thiols	6.79	6.68	-6.68
Point charge	All	Individual	acids	1.54	1.27	1.18
Point charge	All	Individual	amines	0.84	0.65	0.16
Point charge	All	Individual	anilines	0.93	0.87	-0.87
Point charge	All	Individual	heterocycles	1.12	0.76	-0.39
Point charge	All	Individual	indoles	0.72	0.63	0.21
Point charge	All	Individual	phenoles	2.51	2.26	-2.09
Point charge	All	Individual	pyrroles	1.55	1.44	1.44
Point charge	All	Individual	thiols	5.94	5.82	-5.82
Point charge	Individual	All	acids	1.22	0.97	0.53
Point charge	Individual	All	amines	0.84	0.68	0.45
Point charge	Individual	All	anilines	0.39	0.37	-0.03
Point charge	Individual	All	heterocycles	1.04	0.83	0.20
Point charge	Individual	All	indoles	1.86	1.71	1.71
Point charge	Individual	All	phenoles	2.89	2.59	-2.52
Point charge	Individual	All	pyrroles	2.52	2.45	2.45
Point charge	Individual	All	thiols	6.39	6.27	-6.27
Point charge	Individual	Individual	acids	1.57	1.27	1.14
Point charge	Individual	Individual	amines	0.74	0.61	0.21
Point charge	Individual	Individual	anilines	0.93	0.86	-0.86
Point charge	Individual	Individual	heterocycles	1.16	0.69	-0.50
Point charge	Individual	Individual	indoles	0.78	0.72	0.44
Point charge	Individual	Individual	phenoles	2.51	2.25	-2.08
Point charge	Individual	Individual	pyrroles	1.40	1.31	1.31
Point charge	Individual	Individual	thiols	5.71	5.60	-5.60

2.3. Thiol-free models

Table 9.: Re-evaluation of the models shown in Table 3, omitting the substance class "thiols" as defined in the Klicic dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the multipole-based ESP on PM6-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (Mult.)	All	All	acids	1.20	0.99	-0.55
Exact (Mult.)	All	All	amines	0.91	0.74	-0.01
Exact (Mult.)	All	All	anilines	1.22	1.05	1.05
Exact (Mult.)	All	All	heterocycles	1.19	1.02	0.02
Exact (Mult.)	All	All	indoles	1.85	1.63	1.63
Exact (Mult.)	All	All	phenoles	1.00	0.86	0.11
Exact (Mult.)	All	All	pyrroles	2.30	2.20	2.20
Exact (Mult.)	All	Individual	acids	0.96	0.79	0.08
Exact (Mult.)	All	Individual	amines	0.96	0.75	0.09
Exact (Mult.)	All	Individual	anilines	0.78	0.67	0.46
Exact (Mult.)	All	Individual	heterocycles	1.33	0.94	-0.56
Exact (Mult.)	All	Individual	indoles	0.80	0.72	0.18
Exact (Mult.)	All	Individual	phenoles	1.07	0.92	-0.37
Exact (Mult.)	All	Individual	pyrroles	1.09	0.93	0.93
Exact (Mult.)	Individual	All	acids	1.47	1.21	-0.04
Exact (Mult.)	Individual	All	amines	1.08	0.93	-0.25
Exact (Mult.)	Individual	All	anilines	0.90	0.75	0.64
Exact (Mult.)	Individual	All	heterocycles	1.35	1.02	-0.57
Exact (Mult.)	Individual	All	indoles	1.56	1.33	1.33
Exact (Mult.)	Individual	All	phenoles	1.08	0.78	0.33
Exact (Mult.)	Individual	All	pyrroles	1.54	1.44	1.44
Exact (Mult.)	Individual	Individual	acids	1.26	1.03	0.17
Exact (Mult.)	Individual	Individual	amines	1.09	0.89	0.09
Exact (Mult.)	Individual	Individual	anilines	0.84	0.69	0.52
Exact (Mult.)	Individual	Individual	heterocycles	1.44	1.03	-0.69
Exact (Mult.)	Individual	Individual	indoles	1.01	0.94	0.68
Exact (Mult.)	Individual	Individual	phenoles	1.32	1.07	-0.83
Exact (Mult.)	Individual	Individual	pyrroles	0.98	0.86	0.86

Table 10.: Re-evaluation of the models shown in Table 4, omitting the substance class "thiols" as defined in the Klicic dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the NDDO-based ESP on PM6-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (NDDO)	All	All	acids	2.30	2.18	2.18
Exact (NDDO)	All	All	amines	2.78	2.64	-2.64
Exact (NDDO)	All	All	anilines	1.50	1.32	-1.32
Exact (NDDO)	All	All	heterocycles	2.05	1.77	-1.77
Exact (NDDO)	All	All	indoles	0.94	0.84	0.17
Exact (NDDO)	All	All	phenoles	1.60	1.46	1.46
Exact (NDDO)	All	All	pyrroles	1.13	0.86	0.81
Exact (NDDO)	All	Individual	acids	0.75	0.61	0.10
Exact (NDDO)	All	Individual	amines	0.94	0.76	0.07
Exact (NDDO)	All	Individual	anilines	0.64	0.56	-0.05
Exact (NDDO)	All	Individual	heterocycles	1.00	0.75	-0.28
Exact (NDDO)	All	Individual	indoles	0.75	0.54	-0.16
Exact (NDDO)	All	Individual	phenoles	0.82	0.68	-0.48
Exact (NDDO)	All	Individual	pyrroles	1.09	0.91	0.91
Exact (NDDO)	Individual	All	acids	0.97	0.80	-0.27
Exact (NDDO)	Individual	All	amines	0.96	0.76	0.03
Exact (NDDO)	Individual	All	anilines	0.76	0.67	0.37
Exact (NDDO)	Individual	All	heterocycles	0.98	0.84	-0.08
Exact (NDDO)	Individual	All	indoles	1.24	1.12	0.97
Exact (NDDO)	Individual	All	phenoles	0.78	0.59	-0.13
Exact (NDDO)	Individual	All	pyrroles	1.66	1.55	1.55
Exact (NDDO)	Individual	Individual	acids	0.82	0.66	0.11
Exact (NDDO)	Individual	Individual	amines	0.98	0.76	0.10
Exact (NDDO)	Individual	Individual	anilines	0.66	0.59	0.01
Exact (NDDO)	Individual	Individual	heterocycles	1.07	0.79	-0.40
Exact (NDDO)	Individual	Individual	indoles	0.72	0.67	0.15
Exact (NDDO)	Individual	Individual	phenoles	0.88	0.78	-0.54
Exact (NDDO)	Individual	Individual	pyrroles	0.98	0.82	0.82

Table 11.: Re-evaluation of the models shown in Table 5, omitting the substance class "thiols" as defined in the Klicic dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the point-charge-based ESP on PM6-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Point charge	All	All	acids	1.28	1.12	-0.32
Point charge	All	All	amines	0.85	0.76	0.29
Point charge	All	All	anilines	0.60	0.49	0.40
Point charge	All	All	heterocycles	1.15	0.85	0.10
Point charge	All	All	indoles	1.24	1.10	0.89
Point charge	All	All	phenoles	1.75	1.49	-1.23
Point charge	All	All	pyrroles	2.07	1.95	1.95
Point charge	All	Individual	acids	1.25	1.02	0.15
Point charge	All	Individual	amines	0.82	0.74	0.10
Point charge	All	Individual	anilines	0.43	0.34	-0.14
Point charge	All	Individual	heterocycles	1.23	0.83	-0.39
Point charge	All	Individual	indoles	0.82	0.72	-0.03
Point charge	All	Individual	phenoles	1.46	1.20	-0.75
Point charge	All	Individual	pyrroles	1.32	1.14	1.14
Point charge	Individual	All	acids	1.39	1.14	-0.12
Point charge	Individual	All	amines	0.82	0.65	0.27
Point charge	Individual	All	anilines	0.49	0.45	0.23
Point charge	Individual	All	heterocycles	1.14	0.76	-0.18
Point charge	Individual	All	indoles	1.15	1.03	0.81
Point charge	Individual	All	phenoles	1.73	1.46	-1.22
Point charge	Individual	All	pyrroles	1.65	1.54	1.54
Point charge	Individual	Individual	acids	1.39	1.08	0.20
Point charge	Individual	Individual	amines	0.79	0.61	0.15
Point charge	Individual	Individual	anilines	0.43	0.37	-0.13
Point charge	Individual	Individual	heterocycles	1.25	0.77	-0.50
Point charge	Individual	Individual	indoles	0.81	0.71	0.23
Point charge	Individual	Individual	phenoles	1.57	1.29	-0.98
Point charge	Individual	Individual	pyrroles	1.16	1.01	1.01

Table 12.: Re-evaluation of the models shown in Table 6, omitting the substance class "thiols" as defined in the Klicić dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the multipole-based ESP on B3LYP-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (Mult.)	All	All	acids	0.98	0.81	-0.37
Exact (Mult.)	All	All	amines	0.99	0.74	0.36
Exact (Mult.)	All	All	anilines	0.58	0.48	0.31
Exact (Mult.)	All	All	heterocycles	1.10	0.93	0.04
Exact (Mult.)	All	All	indoles	1.65	1.51	1.51
Exact (Mult.)	All	All	phenoles	1.93	1.77	-1.46
Exact (Mult.)	All	All	pyrroles	2.28	2.22	2.22
Exact (Mult.)	All	Individual	acids	0.93	0.70	0.18
Exact (Mult.)	All	Individual	amines	0.95	0.70	0.16
Exact (Mult.)	All	Individual	anilines	0.65	0.57	-0.36
Exact (Mult.)	All	Individual	heterocycles	1.26	0.88	-0.55
Exact (Mult.)	All	Individual	indoles	0.74	0.71	0.43
Exact (Mult.)	All	Individual	phenoles	1.54	1.48	-0.87
Exact (Mult.)	All	Individual	pyrroles	1.32	1.24	1.24
Exact (Mult.)	Individual	All	acids	1.15	0.94	0.07
Exact (Mult.)	Individual	All	amines	0.99	0.74	0.23
Exact (Mult.)	Individual	All	anilines	0.57	0.55	-0.12
Exact (Mult.)	Individual	All	heterocycles	1.29	0.95	-0.52
Exact (Mult.)	Individual	All	indoles	1.28	1.13	1.13
Exact (Mult.)	Individual	All	phenoles	1.77	1.60	-1.24
Exact (Mult.)	Individual	All	pyrroles	1.47	1.42	1.42
Exact (Mult.)	Individual	Individual	acids	1.15	0.94	0.24
Exact (Mult.)	Individual	Individual	amines	0.98	0.75	0.18
Exact (Mult.)	Individual	Individual	anilines	0.65	0.61	-0.30
Exact (Mult.)	Individual	Individual	heterocycles	1.38	1.00	-0.68
Exact (Mult.)	Individual	Individual	indoles	1.03	0.88	0.84
Exact (Mult.)	Individual	Individual	phenoles	1.73	1.56	-1.17
Exact (Mult.)	Individual	Individual	pyrroles	1.21	1.15	1.15

Table 13.: Re-evaluation of the models shown in Table 7, omitting the substance class "thiols" as defined in the Klicić dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the NDDO-based ESP on B3LYP-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Exact (NDDO)	All	All	acids	2.58	2.45	2.45
Exact (NDDO)	All	All	amines	2.75	2.60	-2.60
Exact (NDDO)	All	All	anilines	2.08	2.01	-2.01
Exact (NDDO)	All	All	heterocycles	2.08	1.83	-1.83
Exact (NDDO)	All	All	indoles	0.83	0.77	0.26
Exact (NDDO)	All	All	phenoles	0.70	0.52	0.46
Exact (NDDO)	All	All	pyrroles	1.19	0.97	0.97
Exact (NDDO)	All	Individual	acids	0.96	0.79	0.23
Exact (NDDO)	All	Individual	amines	0.94	0.69	0.12
Exact (NDDO)	All	Individual	anilines	0.98	0.85	-0.85
Exact (NDDO)	All	Individual	heterocycles	1.01	0.78	-0.28
Exact (NDDO)	All	Individual	indoles	0.59	0.57	0.16
Exact (NDDO)	All	Individual	phenoles	1.22	1.14	-1.14
Exact (NDDO)	All	Individual	pyrroles	1.37	1.28	1.28
Exact (NDDO)	Individual	All	acids	1.00	0.81	-0.26
Exact (NDDO)	Individual	All	amines	0.96	0.72	0.34
Exact (NDDO)	Individual	All	anilines	0.53	0.42	-0.18
Exact (NDDO)	Individual	All	heterocycles	0.99	0.84	0.14
Exact (NDDO)	Individual	All	indoles	1.48	1.35	1.35
Exact (NDDO)	Individual	All	phenoles	1.72	1.67	-1.67
Exact (NDDO)	Individual	All	pyrroles	2.12	2.06	2.06
Exact (NDDO)	Individual	Individual	acids	1.00	0.80	0.24
Exact (NDDO)	Individual	Individual	amines	0.92	0.69	0.15
Exact (NDDO)	Individual	Individual	anilines	0.95	0.82	-0.80
Exact (NDDO)	Individual	Individual	heterocycles	1.08	0.82	-0.38
Exact (NDDO)	Individual	Individual	indoles	0.68	0.66	0.38
Exact (NDDO)	Individual	Individual	phenoles	1.25	1.18	-1.18
Exact (NDDO)	Individual	Individual	pyrroles	1.27	1.19	1.19

Table 14.: Re-evaluation of the models shown in Table 8, omitting the substance class "thiols" as defined in the Klicić dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory with the point-charge-based ESP on B3LYP-PCM reoptimised structures.

Potential	Charge fit	pK_a fit	class	RMSE	MAE	MSE
Point charge	All	All	acids	1.05	0.88	-0.04
Point charge	All	All	amines	0.96	0.74	0.49
Point charge	All	All	anilines	0.52	0.38	-0.38
Point charge	All	All	heterocycles	1.05	0.80	0.07
Point charge	All	All	indoles	1.08	0.95	0.82
Point charge	All	All	phenoles	3.02	2.71	-2.65
Point charge	All	All	pyrroles	2.10	2.01	2.01
Point charge	All	Individual	acids	1.25	1.00	0.28
Point charge	All	Individual	amines	0.84	0.65	0.16
Point charge	All	Individual	anilines	0.93	0.87	-0.87
Point charge	All	Individual	heterocycles	1.12	0.76	-0.39
Point charge	All	Individual	indoles	0.72	0.63	0.21
Point charge	All	Individual	phenoles	2.06	1.89	-1.36
Point charge	All	Individual	pyrroles	1.55	1.44	1.44
Point charge	Individual	All	acids	1.13	0.90	0.08
Point charge	Individual	All	amines	0.88	0.71	0.52
Point charge	Individual	All	anilines	0.60	0.48	-0.48
Point charge	Individual	All	heterocycles	1.05	0.72	-0.14
Point charge	Individual	All	indoles	1.09	0.94	0.87
Point charge	Individual	All	phenoles	3.09	2.76	-2.73
Point charge	Individual	All	pyrroles	1.80	1.73	1.73
Point charge	Individual	Individual	acids	1.34	1.04	0.32
Point charge	Individual	Individual	amines	0.74	0.61	0.21
Point charge	Individual	Individual	anilines	0.93	0.86	-0.86
Point charge	Individual	Individual	heterocycles	1.16	0.69	-0.50
Point charge	Individual	Individual	indoles	0.78	0.72	0.44
Point charge	Individual	Individual	phenoles	2.18	1.94	-1.54
Point charge	Individual	Individual	pyrroles	1.40	1.31	1.31

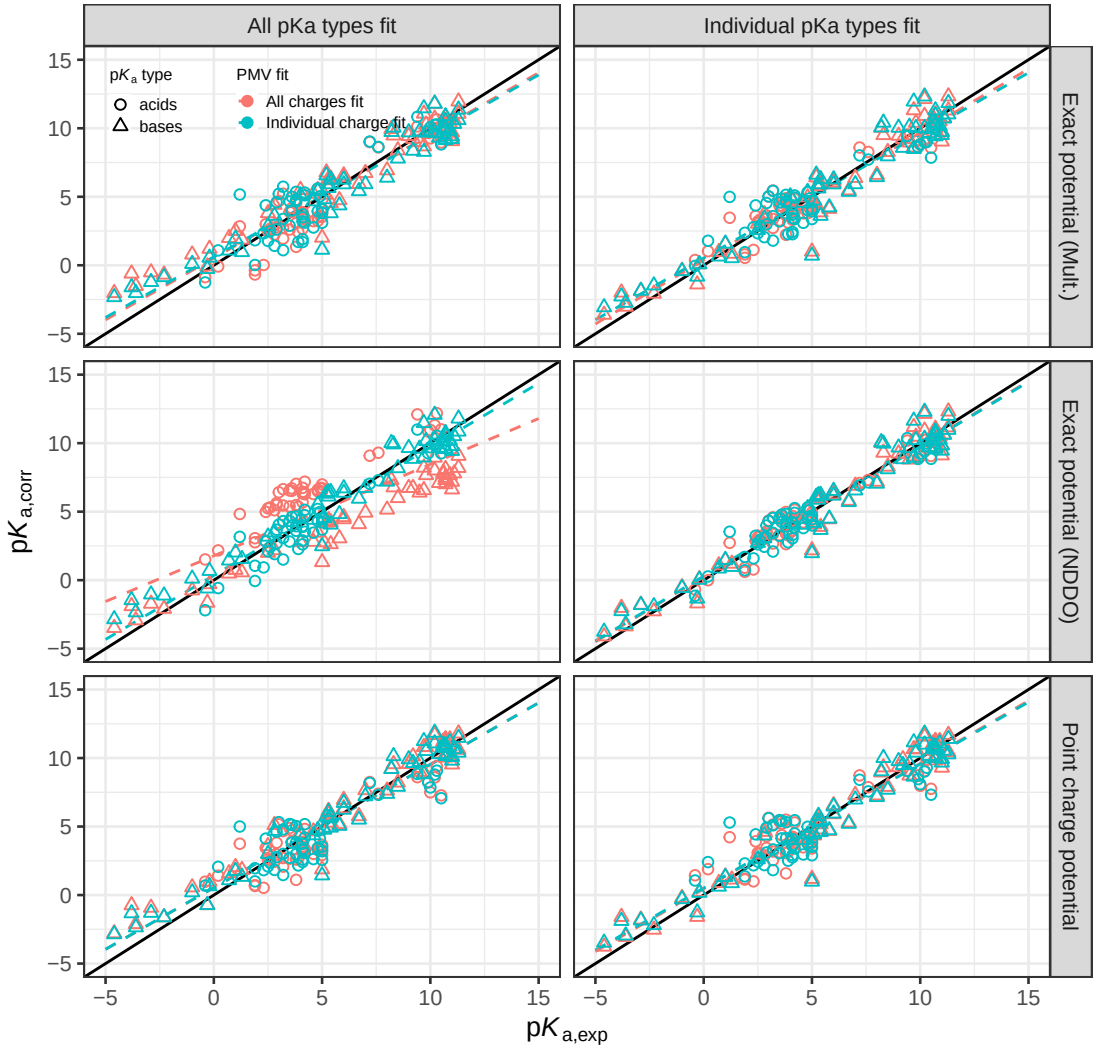

Figure 3.: Re-evaluation of the models presented in figure 5.8 and table 5.10, excluding the substance class "thiols" as defined in the Klicić dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory on PM6-PCM reoptimised structures. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). Table 5.12 shows the corresponding numerical model data.

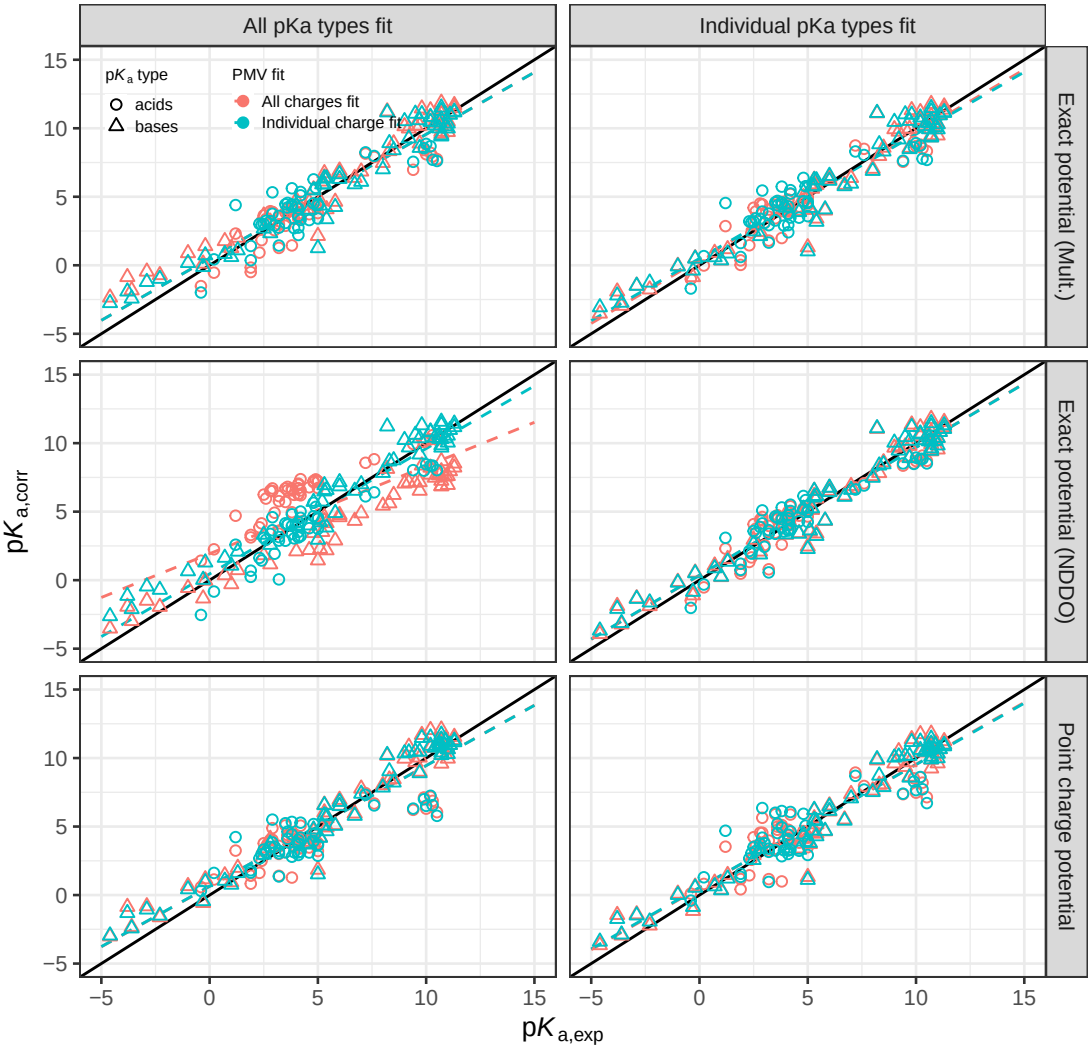

Figure 4.: Re-evaluation of the models presented in figure 5.9 and table 5.11, excluding the substance class "thiols" as defined in the Klicić dataset. Single-point energies were obtained at the ONIOM-EC-RISM/B@PFL level of theory on B3LYP-PCM optimised structures. Circles and triangles represent acids and bases respectively. Colours indicate PMV corrections based on a fit for all charges (red) and individual charge fits (blue). Table 5.13 shows the corresponding numerical model data.

3. Prediction of acidity constants for the SAMPL6 data set

3.1. ONIOM-EC-RISM/X predictions for the PM6-PCM conformers

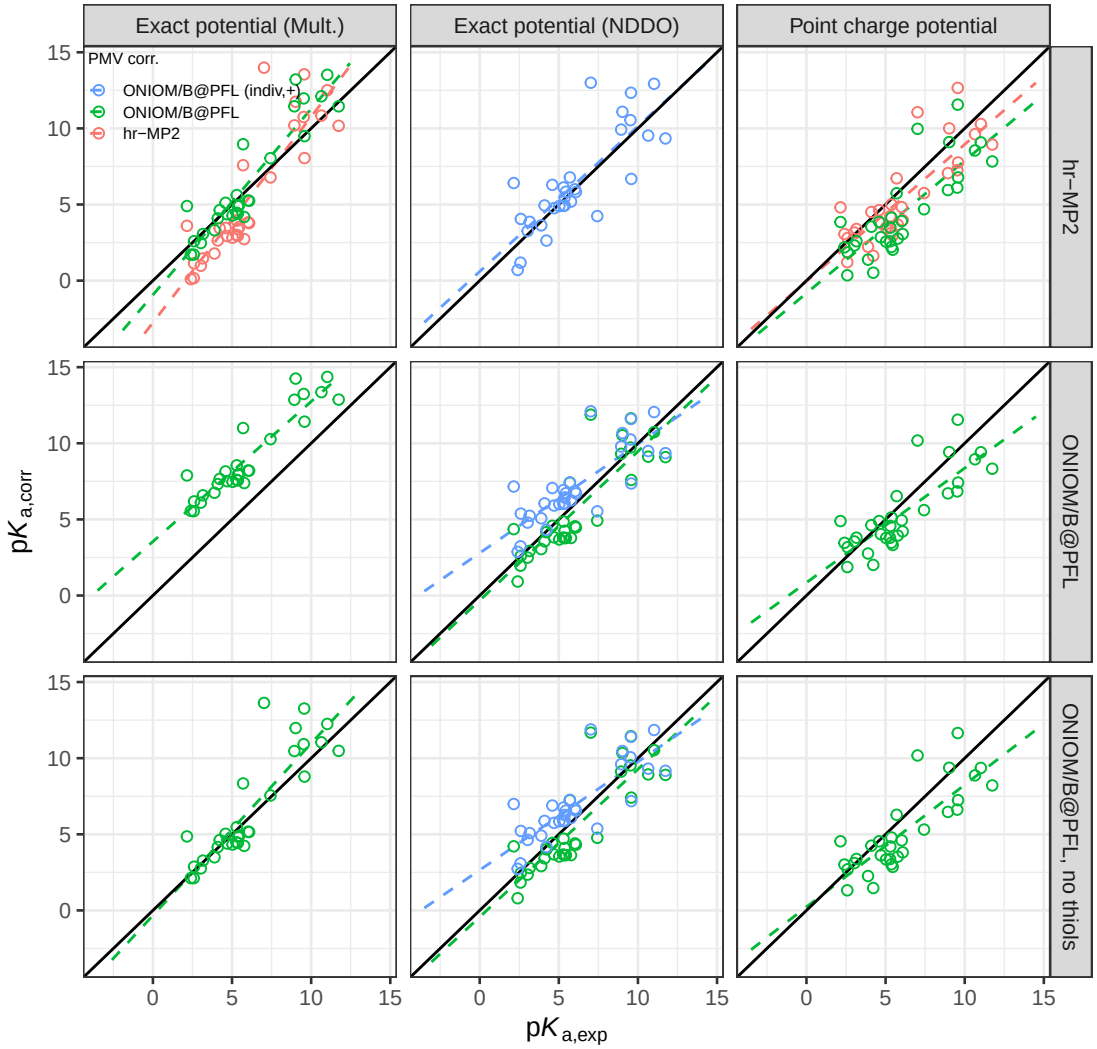

Figure 5.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers reoptimised at the PM6-PCM level of theory. To avoid overplotting, only PMV corrections are shown where the ONIOM addition is performed first. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text.

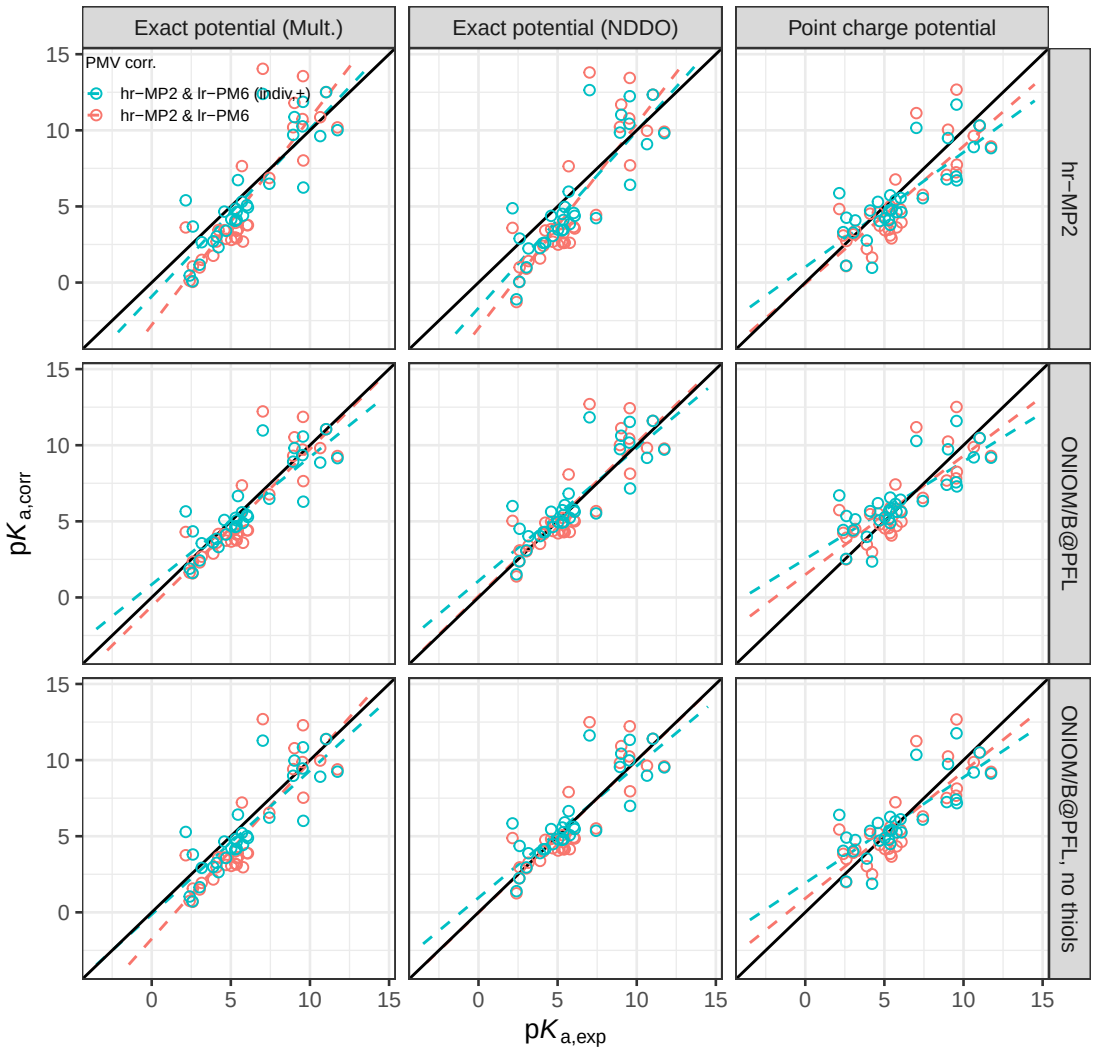

Figure 6.: Predicted versus experimental pK_a values for the SAMPL6 data set. The results are obtained from single point calculations at the ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X level of theory on PCM conformers reoptimised at the PM6-PCM level of theory. To avoid overplotting, only PMV corrections are shown where the ONIOM addition is performed second. The rows show the respective pK_a correction parameters, while the columns represent the ESP supplied to the 3D-RISM solver, as explained in the main text.

Table 15.: Statistical quantities for pK_a prediction on the SAMPL6 dataset. The results were obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/X single point calculations on PM6-PCM-optimised geometries. The best performing *hr*-reference model using the exact potential from the original SAMPL6 publication[24] is shown in the last row. An explanation of the PMV and pK_a fit abbreviations can be found in the main text. See tables 5.6, 5.14, 5.22 and 5.23 for a definition of the model IDs. Models in which thiols were omitted from the training data set are referred to as "n.t." for brevity. The "*hr*-MP2 (All) & *lr*-PM6 (Indiv.,+)" PMV correction is also abbreviated to "*hr*-MP2 & *lr*-PM6 (I+)" for brevity. Please refer to tables 5.14, 5.22 and 5.23 for an overview of where to find the associated PMV and pK_a correction parameters.

Potential	PMV fit	pK_a fit	Model ID	RMSE	MAE	MSE	m'	b'	R^2
Exact (Mult.)	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	2.31	1.99	-0.63	1.35	-2.76	0.77
	/B@PFL (All)	*hr*-MP2	/B|P|M|A‖*hr*|A	2.26	1.43	0.77	1.32	-1.14	0.77
	/B@PFL (All)	/B@PFL	/B|P|M|A‖A	3.64	3.34	3.34	0.99	3.39	0.77
	/B@PFL (All)	/B@PFL, n.t.	/B|P|M|A‖A|nt	1.76	1.14	0.46	1.13	-0.33	0.77
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|P|M|A‖*hr*|A	2.33	2.01	-0.63	1.36	-2.79	0.76
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|P|M|A‖/B|A	1.59	1.25	-0.38	1.03	-0.58	0.76
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|P|M|A‖/B|A|nt	1.94	1.68	-0.75	1.16	-1.75	0.76
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|P|M|I+‖*hr*|A	1.80	1.48	-0.31	1.10	-0.92	0.74
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|P|M|I+‖/B|A	1.41	0.98	-0.14	0.84	0.85	0.74
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|P|M|I+‖/B|A|nt	1.58	1.26	-0.47	0.95	-0.14	0.74
Exact (NDDO)	/B@PFL (All)	/B@PFL	/B|P|N|A‖A	1.63	1.28	-0.45	0.98	-0.32	0.74
	/B@PFL (All)	/B@PFL, n.t.	/B|P|N|A‖A|nt	1.67	1.36	-0.61	0.97	-0.43	0.74
	/B@PFL (Indiv.,+)	*hr*-MP2	/B|P|N|I+‖*hr*|A	1.91	1.38	0.32	0.95	0.60	0.65
	/B@PFL (Indiv.,+)	/B@PFL	/B|P|N|I+‖A	1.93	1.57	1.08	0.72	2.79	0.65
	/B@PFL (Indiv.,+)	/B@PFL, n.t.	/B|P|N|I+‖A|nt	1.84	1.45	0.91	0.71	2.65	0.65
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|P|N|A‖*hr*|A	2.52	2.26	-0.93	1.34	-3.00	0.74
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|P|N|A‖/B|A	1.62	1.17	0.14	1.01	0.09	0.74
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|P|N|A‖/B|A|nt	1.61	1.20	-0.03	1.00	-0.03	0.74
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|P|N|I+‖*hr*|A	2.08	1.78	-0.70	1.16	-1.67	0.73
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|P|N|I+‖/B|A	1.51	1.03	0.31	0.87	1.08	0.73
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|P|N|I+‖/B|A|nt	1.48	1.01	0.15	0.87	0.96	0.73
Point charge	*hr*-MP2 (All)	*hr*-MP2	*hr*|P|E|A‖A	1.80	1.52	-0.65	0.90	-0.04	0.68
	/B@PFL (All)	*hr*-MP2	/B|P|P|A‖*hr*|A	2.33	2.05	-1.61	0.87	-0.82	0.67
	/B@PFL (All)	/B@PFL	/B|P|M|A‖A	1.70	1.47	-0.65	0.75	0.85	0.67
	/B@PFL (All)	/B@PFL, n.t.	/B|P|M|A‖A|nt	1.88	1.59	-0.97	0.80	0.23	0.67
	hr-MP2 & *lr*-PM6 (All)	*hr*-MP2	*hr*&*lr*|P|P|A‖*hr*|A	1.82	1.54	-0.66	0.90	-0.08	0.68
	hr-MP2 & *lr*-PM6 (All)	/B@PFL	*hr*&*lr*|P|P|A‖/B|A	1.58	1.27	0.17	0.78	1.50	0.68
	hr-MP2 & *lr*-PM6 (All)	/B@PFL, n.t.	*hr*&*lr*|P|P|A‖/B|A|nt	1.61	1.35	-0.09	0.83	0.91	0.68
	hr-MP2 & *lr*-PM6 (I+)	*hr*-MP2	*hr*&*lr*|P|P|I+‖*hr*|A	1.73	1.43	-0.47	0.75	1.02	0.64
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL	*hr*&*lr*|P|P|I+‖/B|A	1.67	1.29	0.35	0.64	2.52	0.63
	hr-MP2 & *lr*-PM6 (I+)	/B@PFL, n.t.	*hr*&*lr*|P|P|I+‖/B|A	1.63	1.27	0.08	0.69	1.93	0.64
Exact (*hr*-ref.)[24]	*hr*-MP2 (All)	*hr*-MP2	*hr*|B|E|A‖A	1.13	0.97	-0.36	1.17	-1.38	0.91

3.2. Predicted acidity constants

Table 16.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (Mult.)	SM01	9.53	7.61	8.82	7.67	5.28
Exact (Mult.)	SM02	5.03	2.98	4.58	3.94	3.94
Exact (Mult.)	SM03	7.02	7.14	8.51	5.16	4.41
Exact (Mult.)	SM04	6.02	4.07	5.64	4.99	4.99
Exact (Mult.)	SM05	4.59	3.47	5.28	4.90	4.90
Exact (Mult.)	SM06	3.03	0.80	2.49	1.80	1.80
Exact (Mult.)	SM06	11.74	9.69	11.07	8.15	7.08
Exact (Mult.)	SM07	6.08	4.53	6.07	5.32	5.32
Exact (Mult.)	SM08	4.22	2.52	3.75	0.17	-0.40
Exact (Mult.)	SM09	5.37	2.86	4.44	3.72	3.72
Exact (Mult.)	SM10	9.02	6.70	8.08	4.53	3.93
Exact (Mult.)	SM11	3.89	1.81	3.45	2.33	2.33
Exact (Mult.)	SM12	5.28	3.22	4.80	4.01	4.01
Exact (Mult.)	SM13	5.77	2.81	4.37	3.96	3.96
Exact (Mult.)	SM14	2.58	1.00	2.65	0.28	0.28
Exact (Mult.)	SM14	5.30	3.77	5.56	4.53	4.53
Exact (Mult.)	SM15	4.70	2.93	4.45	3.48	3.48
Exact (Mult.)	SM15	8.94	9.03	10.31	8.81	6.68
Exact (Mult.)	SM16	5.37	4.58	6.22	5.39	5.39
Exact (Mult.)	SM16	10.65	9.70	11.06	8.77	7.23
Exact (Mult.)	SM17	3.16	3.14	4.88	4.19	4.19
Exact (Mult.)	SM18	2.15	1.21	2.86	2.98	2.98
Exact (Mult.)	SM18	9.58	6.45	7.89	1.74	3.09
Exact (Mult.)	SM18	11.02	8.94	10.15	14.10	7.90
Exact (Mult.)	SM19	9.56	7.35	8.76	3.77	4.25
Exact (Mult.)	SM20	5.70	5.00	6.36	1.85	1.97
Exact (Mult.)	SM21	4.10	1.10	2.64	2.30	2.30
Exact (Mult.)	SM22	2.40	0.57	2.25	1.08	1.08
Exact (Mult.)	SM22	7.43	5.85	7.15	6.26	3.67
Exact (Mult.)	SM23	5.45	3.26	4.86	5.25	5.25
Exact (Mult.)	SM24	2.60	1.24	2.83	2.95	2.95

Table 17.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (Mult.)	SM01	9.53	7.70	6.79	4.92
Exact (Mult.)	SM02	5.03	4.37	3.87	3.87
Exact (Mult.)	SM03	7.02	7.45	4.82	4.24
Exact (Mult.)	SM04	6.02	5.20	4.69	4.69
Exact (Mult.)	SM05	4.59	4.92	4.62	4.62
Exact (Mult.)	SM06	3.03	2.73	2.19	2.19
Exact (Mult.)	SM06	11.74	9.46	7.17	6.33
Exact (Mult.)	SM07	6.08	5.54	4.95	4.95
Exact (Mult.)	SM08	4.22	3.72	0.91	0.46
Exact (Mult.)	SM09	5.37	4.26	3.69	3.69
Exact (Mult.)	SM10	9.02	7.11	4.33	3.86
Exact (Mult.)	SM11	3.89	3.48	2.61	2.61
Exact (Mult.)	SM12	5.28	4.54	3.92	3.92
Exact (Mult.)	SM13	5.77	4.21	3.88	3.88
Exact (Mult.)	SM14	2.58	2.85	1.00	1.00
Exact (Mult.)	SM14	5.30	5.14	4.33	4.33
Exact (Mult.)	SM15	4.70	4.26	3.50	3.50
Exact (Mult.)	SM15	8.94	8.86	7.68	6.02
Exact (Mult.)	SM16	5.37	5.66	5.01	5.01
Exact (Mult.)	SM16	10.65	9.45	7.66	6.45
Exact (Mult.)	SM17	3.16	4.60	4.06	4.06
Exact (Mult.)	SM18	2.15	2.48	3.11	3.11
Exact (Mult.)	SM18	9.58	7.21	2.14	3.20
Exact (Mult.)	SM18	11.02	8.99	11.84	6.97
Exact (Mult.)	SM19	9.56	7.65	3.73	4.11
Exact (Mult.)	SM20	5.70	5.76	2.23	2.32
Exact (Mult.)	SM21	4.10	2.85	2.58	2.58
Exact (Mult.)	SM22	2.40	2.54	1.62	1.62
Exact (Mult.)	SM22	7.43	6.38	5.69	3.66
Exact (Mult.)	SM23	5.45	5.29	4.89	4.89
Exact (Mult.)	SM24	2.60	3.60	3.09	3.09

Table 18.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (Mult.)	SM01	9.53	7.66	6.58	4.38
Exact (Mult.)	SM02	5.03	3.72	3.13	3.13
Exact (Mult.)	SM03	7.02	7.36	4.26	3.57
Exact (Mult.)	SM04	6.02	4.71	4.10	4.10
Exact (Mult.)	SM05	4.59	4.37	4.02	4.02
Exact (Mult.)	SM06	3.03	1.79	1.15	1.15
Exact (Mult.)	SM06	11.74	9.74	7.03	6.04
Exact (Mult.)	SM07	6.08	5.11	4.41	4.41
Exact (Mult.)	SM08	4.22	2.96	-0.36	-0.89
Exact (Mult.)	SM09	5.37	3.60	2.92	2.92
Exact (Mult.)	SM10	9.02	6.97	3.68	3.13
Exact (Mult.)	SM11	3.89	2.68	1.64	1.64
Exact (Mult.)	SM12	5.28	3.93	3.19	3.19
Exact (Mult.)	SM13	5.77	3.53	3.15	3.15
Exact (Mult.)	SM14	2.58	1.93	-0.26	-0.26
Exact (Mult.)	SM14	5.30	4.63	3.67	3.67
Exact (Mult.)	SM15	4.70	3.60	2.70	2.70
Exact (Mult.)	SM15	8.94	9.03	7.64	5.67
Exact (Mult.)	SM16	5.37	5.25	4.48	4.48
Exact (Mult.)	SM16	10.65	9.73	7.61	6.18
Exact (Mult.)	SM17	3.16	4.00	3.37	3.37
Exact (Mult.)	SM18	2.15	2.13	2.24	2.24
Exact (Mult.)	SM18	9.58	6.79	1.10	2.34
Exact (Mult.)	SM18	11.02	8.89	12.55	6.80
Exact (Mult.)	SM19	9.56	7.59	2.97	3.42
Exact (Mult.)	SM20	5.70	5.37	1.20	1.31
Exact (Mult.)	SM21	4.10	1.93	1.62	1.62
Exact (Mult.)	SM22	2.40	1.57	0.48	0.48
Exact (Mult.)	SM22	7.43	6.10	5.28	2.89
Exact (Mult.)	SM23	5.45	3.99	4.35	4.35
Exact (Mult.)	SM24	2.60	2.10	2.22	2.22

Table 19.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	11.64	10.42	7.41	12.03
Exact (NDDO)	SM02	5.03	1.44	0.47	4.00	4.00
Exact (NDDO)	SM03	7.02	7.17	6.04	1.97	7.33
Exact (NDDO)	SM04	6.02	2.69	1.72	5.23	5.23
Exact (NDDO)	SM05	4.59	2.02	1.26	4.97	4.97
Exact (NDDO)	SM06	3.03	-0.82	-1.81	1.69	1.69
Exact (NDDO)	SM06	11.74	10.95	9.84	6.02	11.21
Exact (NDDO)	SM07	6.08	3.18	2.18	5.62	5.62
Exact (NDDO)	SM08	4.22	6.50	5.35	1.01	6.56
Exact (NDDO)	SM09	5.37	1.32	0.36	3.82	3.82
Exact (NDDO)	SM10	9.02	7.41	6.29	2.07	7.53
Exact (NDDO)	SM11	3.89	0.34	-0.58	2.64	2.64
Exact (NDDO)	SM12	5.28	1.62	0.63	4.05	4.05
Exact (NDDO)	SM13	5.77	1.41	0.42	4.09	4.09
Exact (NDDO)	SM14	2.58	-1.23	-2.11	0.31	0.31
Exact (NDDO)	SM14	5.30	1.66	0.81	4.10	4.10
Exact (NDDO)	SM15	4.70	1.33	0.30	3.60	3.60
Exact (NDDO)	SM15	8.94	12.43	11.27	8.11	12.84
Exact (NDDO)	SM16	5.37	3.21	2.28	5.68	5.68
Exact (NDDO)	SM16	10.65	10.37	9.24	5.72	10.70
Exact (NDDO)	SM17	3.16	1.45	0.60	4.10	4.10
Exact (NDDO)	SM18	2.15	0.26	-0.64	3.39	3.39
Exact (NDDO)	SM18	9.58	7.30	6.19	0.59	7.02
Exact (NDDO)	SM18	11.02	10.89	9.65	9.38	12.10
Exact (NDDO)	SM19	9.56	7.83	6.74	1.80	7.77
Exact (NDDO)	SM20	5.70	6.26	5.12	0.43	6.22
Exact (NDDO)	SM21	4.10	-1.07	-2.08	1.60	1.60
Exact (NDDO)	SM22	2.40	-2.56	-3.44	-0.29	-0.29
Exact (NDDO)	SM22	7.43	7.05	5.90	3.19	7.60
Exact (NDDO)	SM23	5.45	1.46	0.49	4.68	4.68
Exact (NDDO)	SM24	2.60	0.48	-0.47	3.54	3.54

Table 20.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.90	7.60	11.13
Exact (NDDO)	SM02	5.03	2.30	5.00	5.00
Exact (NDDO)	SM03	7.02	6.55	3.44	7.54
Exact (NDDO)	SM04	6.02	3.25	5.93	5.93
Exact (NDDO)	SM05	4.59	2.90	5.73	5.73
Exact (NDDO)	SM06	3.03	0.56	3.23	3.23
Exact (NDDO)	SM06	11.74	9.46	6.53	10.50
Exact (NDDO)	SM07	6.08	3.60	6.23	6.23
Exact (NDDO)	SM08	4.22	6.02	2.71	6.95
Exact (NDDO)	SM09	5.37	2.21	4.86	4.86
Exact (NDDO)	SM10	9.02	6.74	3.52	7.69
Exact (NDDO)	SM11	3.89	1.49	3.95	3.95
Exact (NDDO)	SM12	5.28	2.42	5.03	5.03
Exact (NDDO)	SM13	5.77	2.26	5.06	5.06
Exact (NDDO)	SM14	2.58	0.33	2.17	2.17
Exact (NDDO)	SM14	5.30	2.56	5.07	5.07
Exact (NDDO)	SM15	4.70	2.16	4.69	4.69
Exact (NDDO)	SM15	8.94	10.55	8.14	11.75
Exact (NDDO)	SM16	5.37	3.68	6.28	6.28
Exact (NDDO)	SM16	10.65	9.00	6.31	10.11
Exact (NDDO)	SM17	3.16	2.39	5.07	5.07
Exact (NDDO)	SM18	2.15	1.45	4.53	4.53
Exact (NDDO)	SM18	9.58	6.67	2.39	7.30
Exact (NDDO)	SM18	11.02	9.31	9.11	11.18
Exact (NDDO)	SM19	9.56	7.09	3.32	7.87
Exact (NDDO)	SM20	5.70	5.85	2.27	6.69
Exact (NDDO)	SM21	4.10	0.35	3.16	3.16
Exact (NDDO)	SM22	2.40	-0.69	1.72	1.72
Exact (NDDO)	SM22	7.43	6.44	4.37	7.75
Exact (NDDO)	SM23	5.45	2.31	5.52	5.52
Exact (NDDO)	SM24	2.60	1.58	4.65	4.65

Table 21.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.73	7.42	10.96
Exact (NDDO)	SM02	5.03	2.10	4.81	4.81
Exact (NDDO)	SM03	7.02	6.37	3.25	7.36
Exact (NDDO)	SM04	6.02	3.06	5.75	5.75
Exact (NDDO)	SM05	4.59	2.71	5.55	5.55
Exact (NDDO)	SM06	3.03	0.35	3.03	3.03
Exact (NDDO)	SM06	11.74	9.28	6.35	10.33
Exact (NDDO)	SM07	6.08	3.41	6.05	6.05
Exact (NDDO)	SM08	4.22	5.84	2.52	6.76
Exact (NDDO)	SM09	5.37	2.01	4.67	4.67
Exact (NDDO)	SM10	9.02	6.56	3.32	7.51
Exact (NDDO)	SM11	3.89	1.29	3.76	3.76
Exact (NDDO)	SM12	5.28	2.22	4.84	4.84
Exact (NDDO)	SM13	5.77	2.06	4.87	4.87
Exact (NDDO)	SM14	2.58	0.12	1.97	1.97
Exact (NDDO)	SM14	5.30	2.36	4.88	4.88
Exact (NDDO)	SM15	4.70	1.97	4.50	4.50
Exact (NDDO)	SM15	8.94	10.38	7.96	11.58
Exact (NDDO)	SM16	5.37	3.49	6.09	6.09
Exact (NDDO)	SM16	10.65	8.82	6.12	9.94
Exact (NDDO)	SM17	3.16	2.20	4.88	4.88
Exact (NDDO)	SM18	2.15	1.25	4.34	4.34
Exact (NDDO)	SM18	9.58	6.49	2.19	7.12
Exact (NDDO)	SM18	11.02	9.13	8.93	11.01
Exact (NDDO)	SM19	9.56	6.91	3.12	7.69
Exact (NDDO)	SM20	5.70	5.66	2.07	6.51
Exact (NDDO)	SM21	4.10	0.15	2.97	2.97
Exact (NDDO)	SM22	2.40	-0.89	1.52	1.52
Exact (NDDO)	SM22	7.43	6.26	4.18	7.57
Exact (NDDO)	SM23	5.45	2.11	5.33	5.33
Exact (NDDO)	SM24	2.60	1.38	4.46	4.46

Table 22.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Point charge	SM01	9.53	8.43	7.42	6.71	5.13
Point charge	SM02	5.03	3.89	3.20	3.13	3.13
Point charge	SM03	7.02	5.99	5.13	3.19	2.40
Point charge	SM04	6.02	5.05	4.34	4.26	4.26
Point charge	SM05	4.59	4.79	4.23	4.36	4.36
Point charge	SM06	3.03	3.17	2.60	2.49	2.49
Point charge	SM06	11.74	11.75	10.87	9.20	8.24
Point charge	SM07	6.08	5.25	4.52	4.37	4.37
Point charge	SM08	4.22	3.93	2.93	0.91	0.17
Point charge	SM09	5.37	3.65	2.95	2.81	2.81
Point charge	SM10	9.02	6.87	6.00	3.99	3.24
Point charge	SM11	3.89	1.56	0.86	0.45	0.45
Point charge	SM12	5.28	4.03	3.31	3.13	3.13
Point charge	SM13	5.77	3.77	3.04	3.14	3.14
Point charge	SM14	2.58	3.17	2.58	1.17	1.17
Point charge	SM14	5.30	4.81	4.27	3.91	3.91
Point charge	SM15	4.70	4.38	3.65	3.32	3.32
Point charge	SM15	8.94	8.86	7.90	6.99	5.53
Point charge	SM16	5.37	6.17	5.53	5.32	5.32
Point charge	SM16	10.65	9.03	8.17	6.82	5.64
Point charge	SM17	3.16	4.50	3.91	3.82	3.82
Point charge	SM18	2.15	3.25	2.54	3.08	3.08
Point charge	SM18	9.58	7.31	6.47	3.06	3.21
Point charge	SM18	11.02	11.39	10.44	12.44	9.12
Point charge	SM19	9.56	8.23	7.40	4.61	4.36
Point charge	SM20	5.70	8.36	7.47	4.95	4.53
Point charge	SM21	4.10	2.69	1.91	2.09	2.09
Point charge	SM22	2.40	2.69	2.04	1.57	1.57
Point charge	SM22	7.43	9.94	9.00	8.43	6.76
Point charge	SM23	5.45	3.01	2.27	3.00	3.00
Point charge	SM24	2.60	2.47	1.65	2.18	2.18

Table 23.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Point charge	SM01	9.53	7.37	6.74	5.34
Point charge	SM02	5.03	3.65	3.59	3.59
Point charge	SM03	7.02	5.35	3.64	2.94
Point charge	SM04	6.02	4.66	4.58	4.58
Point charge	SM05	4.59	4.55	4.67	4.67
Point charge	SM06	3.03	3.12	3.02	3.02
Point charge	SM06	11.74	10.41	8.94	8.09
Point charge	SM07	6.08	4.81	4.68	4.68
Point charge	SM08	4.22	3.41	1.63	0.98
Point charge	SM09	5.37	3.42	3.30	3.30
Point charge	SM10	9.02	6.11	4.34	3.68
Point charge	SM11	3.89	1.59	1.23	1.23
Point charge	SM12	5.28	3.74	3.58	3.58
Point charge	SM13	5.77	3.51	3.60	3.60
Point charge	SM14	2.58	3.10	1.86	1.86
Point charge	SM14	5.30	4.59	4.27	4.27
Point charge	SM15	4.70	4.04	3.76	3.76
Point charge	SM15	8.94	7.79	6.99	5.71
Point charge	SM16	5.37	5.70	5.51	5.51
Point charge	SM16	10.65	8.02	6.84	5.80
Point charge	SM17	3.16	4.27	4.19	4.19
Point charge	SM18	2.15	3.07	3.54	3.54
Point charge	SM18	9.58	6.53	3.53	3.66
Point charge	SM18	11.02	10.02	11.79	8.86
Point charge	SM19	9.56	7.35	4.89	4.67
Point charge	SM20	5.70	7.41	5.19	4.82
Point charge	SM21	4.10	2.51	2.67	2.67
Point charge	SM22	2.40	2.62	2.21	2.21
Point charge	SM22	7.43	8.76	8.25	6.78
Point charge	SM23	5.45	2.83	3.47	3.47
Point charge	SM24	2.60	2.28	2.75	2.75

Table 24.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	/B@PFL (Indiv.)	/B@PFL (Indiv.,+)
Point charge	SM01	9.53	7.20	6.51	4.98
Point charge	SM02	5.03	3.12	3.05	3.05
Point charge	SM03	7.02	4.99	3.11	2.34
Point charge	SM04	6.02	4.22	4.15	4.15
Point charge	SM05	4.59	4.11	4.24	4.24
Point charge	SM06	3.03	2.54	2.43	2.43
Point charge	SM06	11.74	10.54	8.92	7.99
Point charge	SM07	6.08	4.40	4.25	4.25
Point charge	SM08	4.22	2.86	0.90	0.19
Point charge	SM09	5.37	2.87	2.74	2.74
Point charge	SM10	9.02	5.83	3.88	3.16
Point charge	SM11	3.89	0.86	0.46	0.46
Point charge	SM12	5.28	3.22	3.05	3.05
Point charge	SM13	5.77	2.96	3.06	3.06
Point charge	SM14	2.58	2.51	1.16	1.16
Point charge	SM14	5.30	4.15	3.80	3.80
Point charge	SM15	4.70	3.55	3.24	3.24
Point charge	SM15	8.94	7.66	6.78	5.38
Point charge	SM16	5.37	5.37	5.16	5.16
Point charge	SM16	10.65	7.92	6.62	5.48
Point charge	SM17	3.16	3.80	3.72	3.72
Point charge	SM18	2.15	2.48	3.00	3.00
Point charge	SM18	9.58	6.29	2.98	3.13
Point charge	SM18	11.02	10.12	12.06	8.84
Point charge	SM19	9.56	7.18	4.48	4.24
Point charge	SM20	5.70	7.25	4.81	4.40
Point charge	SM21	4.10	1.87	2.04	2.04
Point charge	SM22	2.40	1.99	1.54	1.54
Point charge	SM22	7.43	8.73	8.17	6.56
Point charge	SM23	5.45	2.22	2.92	2.92
Point charge	SM24	2.60	1.62	2.14	2.14

Table 25.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Exact (Mult.)	SM01	9.53	7.63	8.73
Exact (Mult.)	SM02	5.03	3.15	4.61
Exact (Mult.)	SM03	7.02	7.29	8.54
Exact (Mult.)	SM04	6.02	4.26	5.66
Exact (Mult.)	SM05	4.59	3.28	4.81
Exact (Mult.)	SM06	3.03	0.73	2.26
Exact (Mult.)	SM06	11.74	8.46	9.72
Exact (Mult.)	SM07	6.08	4.06	5.51
Exact (Mult.)	SM08	4.22	2.55	3.65
Exact (Mult.)	SM09	5.37	3.25	4.70
Exact (Mult.)	SM10	9.02	6.35	7.57
Exact (Mult.)	SM11	3.89	1.76	3.29
Exact (Mult.)	SM12	5.28	3.34	4.80
Exact (Mult.)	SM13	5.77	2.43	3.85
Exact (Mult.)	SM14	2.58	0.93	2.42
Exact (Mult.)	SM14	5.30	3.73	5.40
Exact (Mult.)	SM15	4.70	2.99	4.40
Exact (Mult.)	SM15	8.94	8.77	9.92
Exact (Mult.)	SM16	5.37	4.60	6.13
Exact (Mult.)	SM16	10.65	9.64	10.89
Exact (Mult.)	SM17	3.16	3.00	4.64
Exact (Mult.)	SM18	2.15	1.22	2.72
Exact (Mult.)	SM18	9.58	6.41	7.75
Exact (Mult.)	SM18	11.02	9.06	10.14
Exact (Mult.)	SM19	9.56	7.30	8.58
Exact (Mult.)	SM20	5.70	4.70	5.97
Exact (Mult.)	SM21	4.10	1.61	3.03
Exact (Mult.)	SM22	2.40	0.65	2.21
Exact (Mult.)	SM22	7.43	5.93	7.11
Exact (Mult.)	SM23	5.45	2.68	4.08
Exact (Mult.)	SM24	2.60	1.28	2.74

Table 26.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	7.72
Exact (Mult.)	SM02	5.03	4.48
Exact (Mult.)	SM03	7.02	7.56
Exact (Mult.)	SM04	6.02	5.31
Exact (Mult.)	SM05	4.59	4.64
Exact (Mult.)	SM06	3.03	2.64
Exact (Mult.)	SM06	11.74	8.49
Exact (Mult.)	SM07	6.08	5.19
Exact (Mult.)	SM08	4.22	3.73
Exact (Mult.)	SM09	5.37	4.56
Exact (Mult.)	SM10	9.02	6.81
Exact (Mult.)	SM11	3.89	3.45
Exact (Mult.)	SM12	5.28	4.63
Exact (Mult.)	SM13	5.77	3.89
Exact (Mult.)	SM14	2.58	2.76
Exact (Mult.)	SM14	5.30	5.11
Exact (Mult.)	SM15	4.70	4.32
Exact (Mult.)	SM15	8.94	8.65
Exact (Mult.)	SM16	5.37	5.68
Exact (Mult.)	SM16	10.65	9.41
Exact (Mult.)	SM17	3.16	4.51
Exact (Mult.)	SM18	2.15	3.00
Exact (Mult.)	SM18	9.58	6.95
Exact (Mult.)	SM18	11.02	8.82
Exact (Mult.)	SM19	9.56	7.60
Exact (Mult.)	SM20	5.70	5.55
Exact (Mult.)	SM21	4.10	3.24
Exact (Mult.)	SM22	2.40	2.61
Exact (Mult.)	SM22	7.43	6.44
Exact (Mult.)	SM23	5.45	4.07
Exact (Mult.)	SM24	2.60	3.02

Table 27.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Exact (Mult.)	SM01	9.53	7.68
Exact (Mult.)	SM02	5.03	3.86
Exact (Mult.)	SM03	7.02	7.50
Exact (Mult.)	SM04	6.02	4.84
Exact (Mult.)	SM05	4.59	4.05
Exact (Mult.)	SM06	3.03	1.68
Exact (Mult.)	SM06	11.74	8.60
Exact (Mult.)	SM07	6.08	4.70
Exact (Mult.)	SM08	4.22	2.97
Exact (Mult.)	SM09	5.37	3.95
Exact (Mult.)	SM10	9.02	6.61
Exact (Mult.)	SM11	3.89	2.64
Exact (Mult.)	SM12	5.28	4.04
Exact (Mult.)	SM13	5.77	3.16
Exact (Mult.)	SM14	2.58	1.83
Exact (Mult.)	SM14	5.30	4.60
Exact (Mult.)	SM15	4.70	3.66
Exact (Mult.)	SM15	8.94	8.78
Exact (Mult.)	SM16	5.37	5.27
Exact (Mult.)	SM16	10.65	9.68
Exact (Mult.)	SM17	3.16	3.89
Exact (Mult.)	SM18	2.15	2.11
Exact (Mult.)	SM18	9.58	6.77
Exact (Mult.)	SM18	11.02	8.99
Exact (Mult.)	SM19	9.56	7.54
Exact (Mult.)	SM20	5.70	5.12
Exact (Mult.)	SM21	4.10	2.39
Exact (Mult.)	SM22	2.40	1.64
Exact (Mult.)	SM22	7.43	6.18
Exact (Mult.)	SM23	5.45	3.37
Exact (Mult.)	SM24	2.60	2.13

Table 28.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	12.05
Exact (NDDO)	SM02	5.03	4.22
Exact (NDDO)	SM03	7.02	7.36
Exact (NDDO)	SM04	6.02	5.36
Exact (NDDO)	SM05	4.59	4.78
Exact (NDDO)	SM06	3.03	1.78
Exact (NDDO)	SM06	11.74	10.63
Exact (NDDO)	SM07	6.08	5.11
Exact (NDDO)	SM08	4.22	6.55
Exact (NDDO)	SM09	5.37	4.25
Exact (NDDO)	SM10	9.02	7.15
Exact (NDDO)	SM11	3.89	2.62
Exact (NDDO)	SM12	5.28	4.25
Exact (NDDO)	SM13	5.77	3.81
Exact (NDDO)	SM14	2.58	0.20
Exact (NDDO)	SM14	5.30	4.07
Exact (NDDO)	SM15	4.70	3.67
Exact (NDDO)	SM15	8.94	12.64
Exact (NDDO)	SM16	5.37	5.69
Exact (NDDO)	SM16	10.65	10.55
Exact (NDDO)	SM17	3.16	3.98
Exact (NDDO)	SM18	2.15	3.40
Exact (NDDO)	SM18	9.58	7.01
Exact (NDDO)	SM18	11.02	12.20
Exact (NDDO)	SM19	9.56	7.76
Exact (NDDO)	SM20	5.70	5.86
Exact (NDDO)	SM21	4.10	2.08
Exact (NDDO)	SM22	2.40	-0.20
Exact (NDDO)	SM22	7.43	7.53
Exact (NDDO)	SM23	5.45	4.13
Exact (NDDO)	SM24	2.60	3.63

Table 29.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
		pKa_exp	/B@PFL (All)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.92	11.15
Exact (NDDO)	SM02	5.03	2.46	5.16
Exact (NDDO)	SM03	7.02	6.58	7.56
Exact (NDDO)	SM04	6.02	3.36	6.04
Exact (NDDO)	SM05	4.59	2.77	5.59
Exact (NDDO)	SM06	3.03	0.63	3.30
Exact (NDDO)	SM06	11.74	9.01	10.06
Exact (NDDO)	SM07	6.08	3.21	5.84
Exact (NDDO)	SM08	4.22	6.02	6.94
Exact (NDDO)	SM09	5.37	2.54	5.19
Exact (NDDO)	SM10	9.02	6.46	7.40
Exact (NDDO)	SM11	3.89	1.47	3.94
Exact (NDDO)	SM12	5.28	2.58	5.19
Exact (NDDO)	SM13	5.77	2.04	4.85
Exact (NDDO)	SM14	2.58	0.24	2.09
Exact (NDDO)	SM14	5.30	2.54	5.05
Exact (NDDO)	SM15	4.70	2.22	4.74
Exact (NDDO)	SM15	8.94	10.39	11.59
Exact (NDDO)	SM16	5.37	3.69	6.28
Exact (NDDO)	SM16	10.65	8.88	10.00
Exact (NDDO)	SM17	3.16	2.31	4.98
Exact (NDDO)	SM18	2.15	1.45	4.53
Exact (NDDO)	SM18	9.58	6.67	7.29
Exact (NDDO)	SM18	11.02	9.38	11.26
Exact (NDDO)	SM19	9.56	7.08	7.86
Exact (NDDO)	SM20	5.70	5.59	6.41
Exact (NDDO)	SM21	4.10	0.72	3.53
Exact (NDDO)	SM22	2.40	-0.62	1.79
Exact (NDDO)	SM22	7.43	6.39	7.69
Exact (NDDO)	SM23	5.45	1.88	5.09
Exact (NDDO)	SM24	2.60	1.63	4.71

Table 30.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
		pKa_exp	/B@PFL (All)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.75	10.98
Exact (NDDO)	SM02	5.03	2.27	4.98
Exact (NDDO)	SM03	7.02	6.39	7.38
Exact (NDDO)	SM04	6.02	3.17	5.85
Exact (NDDO)	SM05	4.59	2.57	5.41
Exact (NDDO)	SM06	3.03	0.43	3.10
Exact (NDDO)	SM06	11.74	8.84	9.88
Exact (NDDO)	SM07	6.08	3.01	5.66
Exact (NDDO)	SM08	4.22	5.83	6.76
Exact (NDDO)	SM09	5.37	2.34	5.00
Exact (NDDO)	SM10	9.02	6.27	7.22
Exact (NDDO)	SM11	3.89	1.27	3.74
Exact (NDDO)	SM12	5.28	2.38	5.00
Exact (NDDO)	SM13	5.77	1.84	4.66
Exact (NDDO)	SM14	2.58	0.04	1.89
Exact (NDDO)	SM14	5.30	2.34	4.86
Exact (NDDO)	SM15	4.70	2.02	4.55
Exact (NDDO)	SM15	8.94	10.22	11.42
Exact (NDDO)	SM16	5.37	3.50	6.10
Exact (NDDO)	SM16	10.65	8.70	9.83
Exact (NDDO)	SM17	3.16	2.11	4.79
Exact (NDDO)	SM18	2.15	1.25	4.34
Exact (NDDO)	SM18	9.58	6.48	7.11
Exact (NDDO)	SM18	11.02	9.21	11.09
Exact (NDDO)	SM19	9.56	6.89	7.68
Exact (NDDO)	SM20	5.70	5.41	6.23
Exact (NDDO)	SM21	4.10	0.51	3.33
Exact (NDDO)	SM22	2.40	-0.83	1.59
Exact (NDDO)	SM22	7.43	6.20	7.51
Exact (NDDO)	SM23	5.45	1.68	4.90
Exact (NDDO)	SM24	2.60	1.43	4.52

Table 31.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Point charge	SM01	9.53	8.48	7.31
Point charge	SM02	5.03	4.05	3.18
Point charge	SM03	7.02	6.20	5.17
Point charge	SM04	6.02	5.10	4.22
Point charge	SM05	4.59	4.30	3.48
Point charge	SM06	3.03	2.97	2.23
Point charge	SM06	11.74	8.76	7.77
Point charge	SM07	6.08	4.30	3.42
Point charge	SM08	4.22	3.98	2.82
Point charge	SM09	5.37	3.82	2.93
Point charge	SM10	9.02	6.57	5.52
Point charge	SM11	3.89	1.57	0.71
Point charge	SM12	5.28	4.08	3.21
Point charge	SM13	5.77	3.59	2.69
Point charge	SM14	2.58	3.03	2.24
Point charge	SM14	5.30	4.79	4.08
Point charge	SM15	4.70	4.36	3.48
Point charge	SM15	8.94	8.67	7.54
Point charge	SM16	5.37	6.20	5.40
Point charge	SM16	10.65	9.01	7.99
Point charge	SM17	3.16	4.49	3.74
Point charge	SM18	2.15	3.21	2.34
Point charge	SM18	9.58	7.17	6.20
Point charge	SM18	11.02	11.48	10.37
Point charge	SM19	9.56	7.73	6.74
Point charge	SM20	5.70	7.63	6.61
Point charge	SM21	4.10	2.98	2.03
Point charge	SM22	2.40	2.71	1.90
Point charge	SM22	7.43	9.94	8.84
Point charge	SM23	5.45	2.53	1.58
Point charge	SM24	2.60	2.42	1.43

Table 32.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	7.40
Point charge	SM02	5.03	3.78
Point charge	SM03	7.02	5.52
Point charge	SM04	6.02	4.69
Point charge	SM05	4.59	4.04
Point charge	SM06	3.03	2.94
Point charge	SM06	11.74	7.81
Point charge	SM07	6.08	3.99
Point charge	SM08	4.22	3.45
Point charge	SM09	5.37	3.55
Point charge	SM10	9.02	5.83
Point charge	SM11	3.89	1.60
Point charge	SM12	5.28	3.80
Point charge	SM13	5.77	3.34
Point charge	SM14	2.58	2.94
Point charge	SM14	5.30	4.57
Point charge	SM15	4.70	4.03
Point charge	SM15	8.94	7.61
Point charge	SM16	5.37	5.73
Point charge	SM16	10.65	8.01
Point charge	SM17	3.16	4.27
Point charge	SM18	2.15	3.03
Point charge	SM18	9.58	6.43
Point charge	SM18	11.02	10.10
Point charge	SM19	9.56	6.91
Point charge	SM20	5.70	6.79
Point charge	SM21	4.10	2.76
Point charge	SM22	2.40	2.64
Point charge	SM22	7.43	8.76
Point charge	SM23	5.45	2.36
Point charge	SM24	2.60	2.23

Table 33.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.16.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
		pKa_exp	/B@PFL (All)
Point charge	SM01	9.53	7.24
Point charge	SM02	5.03	3.26
Point charge	SM03	7.02	5.18
Point charge	SM04	6.02	4.26
Point charge	SM05	4.59	3.55
Point charge	SM06	3.03	2.34
Point charge	SM06	11.74	7.69
Point charge	SM07	6.08	3.49
Point charge	SM08	4.22	2.90
Point charge	SM09	5.37	3.02
Point charge	SM10	9.02	5.51
Point charge	SM11	3.89	0.87
Point charge	SM12	5.28	3.28
Point charge	SM13	5.77	2.78
Point charge	SM14	2.58	2.35
Point charge	SM14	5.30	4.13
Point charge	SM15	4.70	3.54
Point charge	SM15	8.94	7.46
Point charge	SM16	5.37	5.40
Point charge	SM16	10.65	7.91
Point charge	SM17	3.16	3.80
Point charge	SM18	2.15	2.44
Point charge	SM18	9.58	6.17
Point charge	SM18	11.02	10.20
Point charge	SM19	9.56	6.69
Point charge	SM20	5.70	6.57
Point charge	SM21	4.10	2.14
Point charge	SM22	2.40	2.01
Point charge	SM22	7.43	8.73
Point charge	SM23	5.45	1.70
Point charge	SM24	2.60	1.56

Table 34.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Exact (Mult.)	SM01	9.53	8.78	9.89
Exact (Mult.)	SM02	5.03	2.17	3.61
Exact (Mult.)	SM03	7.02	13.15	14.28
Exact (Mult.)	SM04	6.02	3.24	4.63
Exact (Mult.)	SM05	4.59	3.38	4.97
Exact (Mult.)	SM06	3.03	0.43	1.93
Exact (Mult.)	SM06	11.74	8.86	10.12
Exact (Mult.)	SM07	6.08	3.05	4.47
Exact (Mult.)	SM08	4.22	0.50	1.64
Exact (Mult.)	SM09	5.37	2.36	3.78
Exact (Mult.)	SM10	9.02	8.38	9.75
Exact (Mult.)	SM11	3.89	0.64	2.14
Exact (Mult.)	SM12	5.28	2.40	3.84
Exact (Mult.)	SM13	5.77	1.63	3.04
Exact (Mult.)	SM14	2.58	0.34	1.84
Exact (Mult.)	SM14	5.30	3.54	5.22
Exact (Mult.)	SM15	4.70	2.58	3.95
Exact (Mult.)	SM15	8.94	9.77	10.92
Exact (Mult.)	SM16	5.37	3.86	5.35
Exact (Mult.)	SM16	10.65	9.21	10.45
Exact (Mult.)	SM17	3.16	2.37	3.98
Exact (Mult.)	SM18	2.15	2.50	3.83
Exact (Mult.)	SM18	9.58	6.25	7.57
Exact (Mult.)	SM18	11.02	11.33	12.34
Exact (Mult.)	SM19	9.56	10.75	11.87
Exact (Mult.)	SM20	5.70	5.72	7.01
Exact (Mult.)	SM21	4.10	0.97	2.35
Exact (Mult.)	SM22	2.40	0.13	1.66
Exact (Mult.)	SM22	7.43	5.25	6.45
Exact (Mult.)	SM23	5.45	1.58	3.00
Exact (Mult.)	SM24	2.60	0.93	2.36

Table 35.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Exact (Mult.)	SM01	9.53	9.07
Exact (Mult.)	SM02	5.03	4.29
Exact (Mult.)	SM03	7.02	12.40
Exact (Mult.)	SM04	6.02	5.07
Exact (Mult.)	SM05	4.59	5.32
Exact (Mult.)	SM06	3.03	3.01
Exact (Mult.)	SM06	11.74	9.24
Exact (Mult.)	SM07	6.08	4.95
Exact (Mult.)	SM08	4.22	2.79
Exact (Mult.)	SM09	5.37	4.42
Exact (Mult.)	SM10	9.02	8.96
Exact (Mult.)	SM11	3.89	3.17
Exact (Mult.)	SM12	5.28	4.46
Exact (Mult.)	SM13	5.77	3.86
Exact (Mult.)	SM14	2.58	2.95
Exact (Mult.)	SM14	5.30	5.51
Exact (Mult.)	SM15	4.70	4.55
Exact (Mult.)	SM15	8.94	9.85
Exact (Mult.)	SM16	5.37	5.61
Exact (Mult.)	SM16	10.65	9.49
Exact (Mult.)	SM17	3.16	4.57
Exact (Mult.)	SM18	2.15	4.46
Exact (Mult.)	SM18	9.58	7.30
Exact (Mult.)	SM18	11.02	10.93
Exact (Mult.)	SM19	9.56	10.57
Exact (Mult.)	SM20	5.70	6.87
Exact (Mult.)	SM21	4.10	3.33
Exact (Mult.)	SM22	2.40	2.81
Exact (Mult.)	SM22	7.43	6.45
Exact (Mult.)	SM23	5.45	3.82
Exact (Mult.)	SM24	2.60	3.34

Table 36.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.14
Exact (Mult.)	SM02	5.03	3.74
Exact (Mult.)	SM03	7.02	12.90
Exact (Mult.)	SM04	6.02	4.63
Exact (Mult.)	SM05	4.59	4.91
Exact (Mult.)	SM06	3.03	2.30
Exact (Mult.)	SM06	11.74	9.33
Exact (Mult.)	SM07	6.08	4.49
Exact (Mult.)	SM08	4.22	2.06
Exact (Mult.)	SM09	5.37	3.89
Exact (Mult.)	SM10	9.02	9.02
Exact (Mult.)	SM11	3.89	2.49
Exact (Mult.)	SM12	5.28	3.94
Exact (Mult.)	SM13	5.77	3.26
Exact (Mult.)	SM14	2.58	2.23
Exact (Mult.)	SM14	5.30	5.13
Exact (Mult.)	SM15	4.70	4.04
Exact (Mult.)	SM15	8.94	10.02
Exact (Mult.)	SM16	5.37	5.24
Exact (Mult.)	SM16	10.65	9.62
Exact (Mult.)	SM17	3.16	4.06
Exact (Mult.)	SM18	2.15	3.94
Exact (Mult.)	SM18	9.58	7.14
Exact (Mult.)	SM18	11.02	11.24
Exact (Mult.)	SM19	9.56	10.84
Exact (Mult.)	SM20	5.70	6.66
Exact (Mult.)	SM21	4.10	2.67
Exact (Mult.)	SM22	2.40	2.07
Exact (Mult.)	SM22	7.43	6.19
Exact (Mult.)	SM23	5.45	3.22
Exact (Mult.)	SM24	2.60	2.67

Table 37.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	11.91
Exact (NDDO)	SM02	5.03	3.10
Exact (NDDO)	SM03	7.02	11.67
Exact (NDDO)	SM04	6.02	4.23
Exact (NDDO)	SM05	4.59	4.46
Exact (NDDO)	SM06	3.03	1.15
Exact (NDDO)	SM06	11.74	9.66
Exact (NDDO)	SM07	6.08	3.97
Exact (NDDO)	SM08	4.22	4.44
Exact (NDDO)	SM09	5.37	3.20
Exact (NDDO)	SM10	9.02	7.90
Exact (NDDO)	SM11	3.89	1.34
Exact (NDDO)	SM12	5.28	3.15
Exact (NDDO)	SM13	5.77	2.94
Exact (NDDO)	SM14	2.58	-0.78
Exact (NDDO)	SM14	5.30	3.61
Exact (NDDO)	SM15	4.70	3.04
Exact (NDDO)	SM15	8.94	12.33
Exact (NDDO)	SM16	5.37	4.73
Exact (NDDO)	SM16	10.65	9.51
Exact (NDDO)	SM17	3.16	3.15
Exact (NDDO)	SM18	2.15	4.50
Exact (NDDO)	SM18	9.58	6.30
Exact (NDDO)	SM18	11.02	11.59
Exact (NDDO)	SM19	9.56	8.86
Exact (NDDO)	SM20	5.70	5.80
Exact (NDDO)	SM21	4.10	1.45
Exact (NDDO)	SM22	2.40	-1.01
Exact (NDDO)	SM22	7.43	5.89
Exact (NDDO)	SM23	5.45	3.14
Exact (NDDO)	SM24	2.60	3.10

Table 38.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$ pKa_exp	p$K_{a,corr}$ /B@PFL (All)	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.78	11.29
Exact (NDDO)	SM02	5.03	2.34	4.67
Exact (NDDO)	SM03	7.02	10.89	11.10
Exact (NDDO)	SM04	6.02	3.21	5.52
Exact (NDDO)	SM05	4.59	3.20	5.69
Exact (NDDO)	SM06	3.03	0.91	3.21
Exact (NDDO)	SM06	11.74	9.32	9.59
Exact (NDDO)	SM07	6.08	3.07	5.32
Exact (NDDO)	SM08	4.22	5.55	5.67
Exact (NDDO)	SM09	5.37	2.47	4.74
Exact (NDDO)	SM10	9.02	8.14	8.28
Exact (NDDO)	SM11	3.89	1.30	3.34
Exact (NDDO)	SM12	5.28	2.48	4.70
Exact (NDDO)	SM13	5.77	2.08	4.55
Exact (NDDO)	SM14	2.58	0.48	1.75
Exact (NDDO)	SM14	5.30	2.95	5.05
Exact (NDDO)	SM15	4.70	2.51	4.62
Exact (NDDO)	SM15	8.94	11.13	11.60
Exact (NDDO)	SM16	5.37	3.69	5.90
Exact (NDDO)	SM16	10.65	9.11	9.48
Exact (NDDO)	SM17	3.16	2.41	4.71
Exact (NDDO)	SM18	2.15	2.90	5.72
Exact (NDDO)	SM18	9.58	7.30	7.07
Exact (NDDO)	SM18	11.02	9.74	11.04
Exact (NDDO)	SM19	9.56	9.03	8.99
Exact (NDDO)	SM20	5.70	6.69	6.70
Exact (NDDO)	SM21	4.10	0.95	3.43
Exact (NDDO)	SM22	2.40	-0.38	1.58
Exact (NDDO)	SM22	7.43	6.16	6.76
Exact (NDDO)	SM23	5.45	1.73	4.70
Exact (NDDO)	SM24	2.60	1.87	4.67

Table 39.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			/B@PFL (All)	/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.59	11.09
Exact (NDDO)	SM02	5.03	2.20	4.52
Exact (NDDO)	SM03	7.02	10.70	10.91
Exact (NDDO)	SM04	6.02	3.07	5.36
Exact (NDDO)	SM05	4.59	3.05	5.53
Exact (NDDO)	SM06	3.03	0.79	3.07
Exact (NDDO)	SM06	11.74	9.14	9.41
Exact (NDDO)	SM07	6.08	2.93	5.17
Exact (NDDO)	SM08	4.22	5.39	5.51
Exact (NDDO)	SM09	5.37	2.33	4.59
Exact (NDDO)	SM10	9.02	7.96	8.10
Exact (NDDO)	SM11	3.89	1.17	3.20
Exact (NDDO)	SM12	5.28	2.34	4.55
Exact (NDDO)	SM13	5.77	1.95	4.40
Exact (NDDO)	SM14	2.58	0.36	1.62
Exact (NDDO)	SM14	5.30	2.82	4.89
Exact (NDDO)	SM15	4.70	2.37	4.47
Exact (NDDO)	SM15	8.94	10.93	11.40
Exact (NDDO)	SM16	5.37	3.55	5.73
Exact (NDDO)	SM16	10.65	8.93	9.29
Exact (NDDO)	SM17	3.16	2.27	4.55
Exact (NDDO)	SM18	2.15	2.76	5.56
Exact (NDDO)	SM18	9.58	7.13	6.90
Exact (NDDO)	SM18	11.02	9.55	10.85
Exact (NDDO)	SM19	9.56	8.84	8.81
Exact (NDDO)	SM20	5.70	6.52	6.53
Exact (NDDO)	SM21	4.10	0.82	3.29
Exact (NDDO)	SM22	2.40	-0.50	1.45
Exact (NDDO)	SM22	7.43	6.00	6.60
Exact (NDDO)	SM23	5.45	1.60	4.55
Exact (NDDO)	SM24	2.60	1.74	4.51

Table 40.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Point charge	SM01	9.53	9.41	8.25
Point charge	SM02	5.03	3.27	2.38
Point charge	SM03	7.02	12.16	11.07
Point charge	SM04	6.02	4.46	3.56
Point charge	SM05	4.59	5.07	4.31
Point charge	SM06	3.03	2.67	1.92
Point charge	SM06	11.74	11.19	10.17
Point charge	SM07	6.08	3.66	2.78
Point charge	SM08	4.22	1.99	0.87
Point charge	SM09	5.37	3.10	2.19
Point charge	SM10	9.02	8.89	7.94
Point charge	SM11	3.89	0.87	0.01
Point charge	SM12	5.28	3.35	2.46
Point charge	SM13	5.77	3.05	2.14
Point charge	SM14	2.58	2.55	1.74
Point charge	SM14	5.30	4.69	4.02
Point charge	SM15	4.70	4.06	3.15
Point charge	SM15	8.94	9.41	8.28
Point charge	SM16	5.37	5.55	4.72
Point charge	SM16	10.65	8.65	7.63
Point charge	SM17	3.16	3.91	3.14
Point charge	SM18	2.15	7.00	6.10
Point charge	SM18	9.58	5.62	4.67
Point charge	SM18	11.02	12.84	11.65
Point charge	SM19	9.56	10.61	9.48
Point charge	SM20	5.70	8.40	7.39
Point charge	SM21	4.10	2.49	1.51
Point charge	SM22	2.40	2.43	1.60
Point charge	SM22	7.43	9.07	7.99
Point charge	SM23	5.45	2.01	1.07
Point charge	SM24	2.60	2.66	1.65

Table 41.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	8.69
Point charge	SM02	5.03	3.62
Point charge	SM03	7.02	11.13
Point charge	SM04	6.02	4.64
Point charge	SM05	4.59	5.29
Point charge	SM06	3.03	3.22
Point charge	SM06	11.74	10.35
Point charge	SM07	6.08	3.96
Point charge	SM08	4.22	2.31
Point charge	SM09	5.37	3.46
Point charge	SM10	9.02	8.42
Point charge	SM11	3.89	1.57
Point charge	SM12	5.28	3.68
Point charge	SM13	5.77	3.41
Point charge	SM14	2.58	3.07
Point charge	SM14	5.30	5.04
Point charge	SM15	4.70	4.28
Point charge	SM15	8.94	8.72
Point charge	SM16	5.37	5.64
Point charge	SM16	10.65	8.16
Point charge	SM17	3.16	4.27
Point charge	SM18	2.15	6.83
Point charge	SM18	9.58	5.60
Point charge	SM18	11.02	11.63
Point charge	SM19	9.56	9.76
Point charge	SM20	5.70	7.94
Point charge	SM21	4.10	2.86
Point charge	SM22	2.40	2.94
Point charge	SM22	7.43	8.47
Point charge	SM23	5.45	2.48
Point charge	SM24	2.60	2.99

Table 42.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.17.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	8.59
Point charge	SM02	5.03	3.19
Point charge	SM03	7.02	11.19
Point charge	SM04	6.02	4.27
Point charge	SM05	4.59	4.97
Point charge	SM06	3.03	2.76
Point charge	SM06	11.74	10.37
Point charge	SM07	6.08	3.55
Point charge	SM08	4.22	1.79
Point charge	SM09	5.37	3.01
Point charge	SM10	9.02	8.31
Point charge	SM11	3.89	1.00
Point charge	SM12	5.28	3.25
Point charge	SM13	5.77	2.97
Point charge	SM14	2.58	2.60
Point charge	SM14	5.30	4.70
Point charge	SM15	4.70	3.89
Point charge	SM15	8.94	8.62
Point charge	SM16	5.37	5.34
Point charge	SM16	10.65	8.02
Point charge	SM17	3.16	3.88
Point charge	SM18	2.15	6.61
Point charge	SM18	9.58	5.29
Point charge	SM18	11.02	11.73
Point charge	SM19	9.56	9.73
Point charge	SM20	5.70	7.80
Point charge	SM21	4.10	2.38
Point charge	SM22	2.40	2.46
Point charge	SM22	7.43	8.35
Point charge	SM23	5.45	1.97
Point charge	SM24	2.60	2.51

Table 43.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/B@PFL level of theory using the multipole-based ESP and /B@PFL PMV correction parameters from Table 5.5 with one set of parameters for all charge states ("/B@PFL (All)") on the B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of pK_a correction parameters used. These values correspond to the results shown in table 5.18.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2	/B@PFL
Exact (Mult.)	SM01	9.53	9.31	8.08
Exact (Mult.)	SM02	5.03	4.80	4.54
Exact (Mult.)	SM03	7.02	7.63	6.76
Exact (Mult.)	SM04	6.02	5.67	5.22
Exact (Mult.)	SM05	4.59	8.26	7.26
Exact (Mult.)	SM06	3.03	2.39	2.65
Exact (Mult.)	SM06	11.74	11.47	9.78
Exact (Mult.)	SM07	6.08	6.17	5.62
Exact (Mult.)	SM08	4.22	4.10	3.99
Exact (Mult.)	SM09	5.37	5.44	5.04
Exact (Mult.)	SM10	9.02	9.13	7.94
Exact (Mult.)	SM11	3.89	3.53	3.55
Exact (Mult.)	SM12	5.28	5.26	4.90
Exact (Mult.)	SM13	5.77	5.94	5.44
Exact (Mult.)	SM14	2.58	2.91	3.06
Exact (Mult.)	SM14	5.30	5.67	5.22
Exact (Mult.)	SM15	4.70	4.76	4.51
Exact (Mult.)	SM15	8.94	10.13	8.73
Exact (Mult.)	SM16	5.37	7.04	6.29
Exact (Mult.)	SM16	10.65	9.88	8.52
Exact (Mult.)	SM17	3.16	5.75	5.29
Exact (Mult.)	SM18	2.15	2.54	2.77
Exact (Mult.)	SM18	9.58	8.65	7.56
Exact (Mult.)	SM18	11.02	11.31	9.65
Exact (Mult.)	SM19	9.56	8.44	7.40
Exact (Mult.)	SM20	5.70	7.62	6.76
Exact (Mult.)	SM21	4.10	2.85	3.01
Exact (Mult.)	SM22	2.40	2.76	2.94
Exact (Mult.)	SM22	7.43	8.46	7.41
Exact (Mult.)	SM23	5.45	5.72	5.26
Exact (Mult.)	SM24	2.60	4.46	4.28

Table 44.: Differences between the pK_a values obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B and /B@PFL calculations on the B3LYP-PCM-optimised structures and the /B@PFL PMV parameters from Table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"), as well as the *hr* pK_a correction parameters. The values are categorised by ONIOM partitions to allow estimation of the partitioning effect on the predicted pK_a values.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}^{/B}$ − p$K_{a,corr}^{/B@PFL}$			
			P1	P2	P3	P4
Exact (Mult.)	SM01	9.53	-0.49			
Exact (Mult.)	SM02	5.03	-0.22			
Exact (Mult.)	SM03	7.02	0.88			
Exact (Mult.)	SM04	6.02	-0.03			
Exact (Mult.)	SM05	4.59	-2.98			
Exact (Mult.)	SM06	3.03	0.10			
Exact (Mult.)	SM06	11.74	-0.40			
Exact (Mult.)	SM07	6.08	-0.10			
Exact (Mult.)	SM08	4.22	-0.35			
Exact (Mult.)	SM09	5.37	-1.00			
Exact (Mult.)	SM10	9.02	-1.05			
Exact (Mult.)	SM11	3.89	-0.08			
Exact (Mult.)	SM12	5.28	-0.46			
Exact (Mult.)	SM13	5.77	-1.57			
Exact (Mult.)	SM14	2.58	-0.26			
Exact (Mult.)	SM14	5.30	-0.11			
Exact (Mult.)	SM15	4.70	-0.31			
Exact (Mult.)	SM15	8.94	0.18			
Exact (Mult.)	SM16	5.37	-0.82			
Exact (Mult.)	SM16	10.65	1.18			
Exact (Mult.)	SM17	3.16	-0.87			
Exact (Mult.)	SM18	2.15	-0.37	-0.38	-0.29	0.32
Exact (Mult.)	SM18	9.58	-0.45	-0.49	-0.33	-0.76
Exact (Mult.)	SM18	11.02	-0.84	-1.04	-1.14	-1.16
Exact (Mult.)	SM19	9.56	0.32			
Exact (Mult.)	SM20	5.70	-1.26			
Exact (Mult.)	SM21	4.10	-0.21			
Exact (Mult.)	SM22	2.40	-0.51			
Exact (Mult.)	SM22	7.43	-1.31			
Exact (Mult.)	SM23	5.45	0.03	0.86	-0.22	-0.86
Exact (Mult.)	SM24	2.60	-0.86	-0.46	-0.77	-1.63

Table 45.: Differences between the pK_a values obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B and /B@PFL calculations on the B3LYP-PCM-optimised structures and the /B@PFL PMV parameters from Table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"), as well as the /B@PFL pK_a correction parameters. The values are categorised by ONIOM partitions to allow estimation of the partitioning effect on the predicted pK_a values.

Potential	Name	p$K_{a,exp}$	$pK_{a,corr}^{/B} - pK_{a,corr}^{/B@PFL}$			
			P1	P2	P3	P4
Exact (Mult.)	SM01	9.53	-0.38			
Exact (Mult.)	SM02	5.03	-0.17			
Exact (Mult.)	SM03	7.02	0.69			
Exact (Mult.)	SM04	6.02	-0.02			
Exact (Mult.)	SM05	4.59	-2.34			
Exact (Mult.)	SM06	3.03	0.08			
Exact (Mult.)	SM06	11.74	-0.32			
Exact (Mult.)	SM07	6.08	-0.08			
Exact (Mult.)	SM08	4.22	-0.27			
Exact (Mult.)	SM09	5.37	-0.78			
Exact (Mult.)	SM10	9.02	-0.83			
Exact (Mult.)	SM11	3.89	-0.07			
Exact (Mult.)	SM12	5.28	-0.36			
Exact (Mult.)	SM13	5.77	-1.23			
Exact (Mult.)	SM14	2.58	-0.21			
Exact (Mult.)	SM14	5.30	-0.08			
Exact (Mult.)	SM15	4.70	-0.25			
Exact (Mult.)	SM15	8.94	0.13			
Exact (Mult.)	SM16	5.37	-0.63			
Exact (Mult.)	SM16	10.65	0.93			
Exact (Mult.)	SM17	3.16	-0.69			
Exact (Mult.)	SM18	2.15	-0.29	-0.30	-0.23	-0.29
Exact (Mult.)	SM18	9.58	-0.35	-0.38	-0.26	-0.35
Exact (Mult.)	SM18	11.02	-0.66	-0.82	-0.89	-0.66
Exact (Mult.)	SM19	9.56	0.25			
Exact (Mult.)	SM20	5.70	-1.00			
Exact (Mult.)	SM21	4.10	-0.16			
Exact (Mult.)	SM22	2.40	-0.40			
Exact (Mult.)	SM22	7.43	-1.03			
Exact (Mult.)	SM23	5.45	0.03	0.68	-0.17	0.03
Exact (Mult.)	SM24	2.60	-0.68	-0.36	-0.61	-0.68

Table 46.: Differences between experimental and predicted pK_a values obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B calculations on the B3LYP-PCM-optimised structures and the /B@PFL PMV parameters from Table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"), as well as the *hr* pK_a correction parameters. The values are categorised by ONIOM partitions to allow estimation of the partitioning effect on the predicted pK_a values.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}^{/B}$ − p$K_{a,exp}$			
			P1	P2	P3	P4
Exact (Mult.)	SM01	9.53	-0.71			
Exact (Mult.)	SM02	5.03	-0.45			
Exact (Mult.)	SM03	7.02	1.49			
Exact (Mult.)	SM04	6.02	-0.38			
Exact (Mult.)	SM05	4.59	0.69			
Exact (Mult.)	SM06	3.03	-0.54			
Exact (Mult.)	SM06	11.74	-0.67			
Exact (Mult.)	SM07	6.08	-0.01			
Exact (Mult.)	SM08	4.22	-0.47			
Exact (Mult.)	SM09	5.37	-0.93			
Exact (Mult.)	SM10	9.02	-0.94			
Exact (Mult.)	SM11	3.89	-0.44			
Exact (Mult.)	SM12	5.28	-0.48			
Exact (Mult.)	SM13	5.77	-1.40			
Exact (Mult.)	SM14	2.58	0.07			
Exact (Mult.)	SM14	5.30	0.26			
Exact (Mult.)	SM15	4.70	-0.25			
Exact (Mult.)	SM15	8.94	1.37			
Exact (Mult.)	SM16	5.37	0.85			
Exact (Mult.)	SM16	10.65	0.41			
Exact (Mult.)	SM17	3.16	1.72			
Exact (Mult.)	SM18	2.15	0.02	0.01	0.10	0.71
Exact (Mult.)	SM18	9.58	-1.38	-1.42	-1.26	-1.69
Exact (Mult.)	SM18	11.02	-0.55	-0.75	-0.85	-0.87
Exact (Mult.)	SM19	9.56	-0.80			
Exact (Mult.)	SM20	5.70	0.66			
Exact (Mult.)	SM21	4.10	-1.46			
Exact (Mult.)	SM22	2.40	-0.15			
Exact (Mult.)	SM22	7.43	-0.28			
Exact (Mult.)	SM23	5.45	0.30	1.13	0.05	-0.59
Exact (Mult.)	SM24	2.60	1.00	1.40	1.09	0.23

Table 47.: Differences between experimental and predicted pK_a values obtained from ONIOM(MP2/6-311+G**:PM6)-EC-RISM/B calculations on the B3LYP-PCM-optimised structures and the /B@PFL PMV parameters from Table 5.5 with one set of parameters for all charge states ("/B@PFL (All)"), as well as the /B@PFL pK_a correction parameters. The values are categorised by ONIOM partitions to allow estimation of the partitioning effect on the predicted pK_a values.

Potential	Name	p$K_{a,exp}$	p$K^{/B}_{a,corr}$ − p$K_{a,exp}$			
			P1	P2	P3	P4
Exact (Mult.)	SM01	9.53	-1.83			
Exact (Mult.)	SM02	5.03	-0.66			
Exact (Mult.)	SM03	7.02	0.43			
Exact (Mult.)	SM04	6.02	-0.82			
Exact (Mult.)	SM05	4.59	0.33			
Exact (Mult.)	SM06	3.03	-0.30			
Exact (Mult.)	SM06	11.74	-2.28			
Exact (Mult.)	SM07	6.08	-0.54			
Exact (Mult.)	SM08	4.22	-0.50			
Exact (Mult.)	SM09	5.37	-1.11			
Exact (Mult.)	SM10	9.02	-1.91			
Exact (Mult.)	SM11	3.89	-0.41			
Exact (Mult.)	SM12	5.28	-0.74			
Exact (Mult.)	SM13	5.77	-1.56			
Exact (Mult.)	SM14	2.58	0.27			
Exact (Mult.)	SM14	5.30	-0.16			
Exact (Mult.)	SM15	4.70	-0.44			
Exact (Mult.)	SM15	8.94	-0.08			
Exact (Mult.)	SM16	5.37	0.29			
Exact (Mult.)	SM16	10.65	-1.20			
Exact (Mult.)	SM17	3.16	1.44			
Exact (Mult.)	SM18	2.15	0.33	0.32	0.39	0.33
Exact (Mult.)	SM18	9.58	-2.37	-2.40	-2.28	-2.37
Exact (Mult.)	SM18	11.02	-2.03	-2.19	-2.26	-2.03
Exact (Mult.)	SM19	9.56	-1.91			
Exact (Mult.)	SM20	5.70	0.06			
Exact (Mult.)	SM21	4.10	-1.25			
Exact (Mult.)	SM22	2.40	0.14			
Exact (Mult.)	SM22	7.43	-1.05			
Exact (Mult.)	SM23	5.45	-0.16	0.49	-0.36	-0.16
Exact (Mult.)	SM24	2.60	1.00	1.32	1.07	1.00

Table 48.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.48	10.78
Exact (Mult.)	SM02	5.03	3.44	5.07
Exact (Mult.)	SM03	7.02	8.90	10.32
Exact (Mult.)	SM04	6.02	4.31	5.93
Exact (Mult.)	SM05	4.59	2.87	4.69
Exact (Mult.)	SM06	3.03	1.07	2.69
Exact (Mult.)	SM06	11.74	11.28	12.65
Exact (Mult.)	SM07	6.08	4.74	6.32
Exact (Mult.)	SM08	4.22	5.15	6.42
Exact (Mult.)	SM09	5.37	3.42	5.05
Exact (Mult.)	SM10	9.02	9.88	11.35
Exact (Mult.)	SM11	3.89	2.77	4.44
Exact (Mult.)	SM12	5.28	3.58	5.18
Exact (Mult.)	SM13	5.77	3.87	5.47
Exact (Mult.)	SM14	2.58	0.54	2.31
Exact (Mult.)	SM14	5.30	4.08	5.78
Exact (Mult.)	SM15	4.70	3.13	4.68
Exact (Mult.)	SM15	8.94	9.64	10.99
Exact (Mult.)	SM16	5.37	4.17	5.82
Exact (Mult.)	SM16	10.65	11.53	12.94
Exact (Mult.)	SM17	3.16	2.35	4.07
Exact (Mult.)	SM18	2.15	1.65	3.35
Exact (Mult.)	SM18	9.58	8.78	10.26
Exact (Mult.)	SM18	11.02	10.44	11.63
Exact (Mult.)	SM19	9.56	10.31	11.77
Exact (Mult.)	SM20	5.70	7.15	8.59
Exact (Mult.)	SM21	4.10	2.93	4.53
Exact (Mult.)	SM22	2.40	-0.13	1.57
Exact (Mult.)	SM22	7.43	5.76	7.08
Exact (Mult.)	SM23	5.45	4.29	5.95
Exact (Mult.)	SM24	2.60	1.21	2.82

Table 49.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Exact (Mult.)	SM01	9.53	9.23
Exact (Mult.)	SM02	5.03	4.75
Exact (Mult.)	SM03	7.02	8.87
Exact (Mult.)	SM04	6.02	5.42
Exact (Mult.)	SM05	4.59	4.45
Exact (Mult.)	SM06	3.03	2.89
Exact (Mult.)	SM06	11.74	10.70
Exact (Mult.)	SM07	6.08	5.73
Exact (Mult.)	SM08	4.22	5.82
Exact (Mult.)	SM09	5.37	4.74
Exact (Mult.)	SM10	9.02	9.68
Exact (Mult.)	SM11	3.89	4.25
Exact (Mult.)	SM12	5.28	4.84
Exact (Mult.)	SM13	5.77	5.07
Exact (Mult.)	SM14	2.58	2.59
Exact (Mult.)	SM14	5.30	5.31
Exact (Mult.)	SM15	4.70	4.45
Exact (Mult.)	SM15	8.94	9.40
Exact (Mult.)	SM16	5.37	5.34
Exact (Mult.)	SM16	10.65	10.92
Exact (Mult.)	SM17	3.16	3.97
Exact (Mult.)	SM18	2.15	3.40
Exact (Mult.)	SM18	9.58	8.83
Exact (Mult.)	SM18	11.02	9.90
Exact (Mult.)	SM19	9.56	10.01
Exact (Mult.)	SM20	5.70	7.51
Exact (Mult.)	SM21	4.10	4.33
Exact (Mult.)	SM22	2.40	2.01
Exact (Mult.)	SM22	7.43	6.33
Exact (Mult.)	SM23	5.45	5.45
Exact (Mult.)	SM24	2.60	2.98

Table 50.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.47
Exact (Mult.)	SM02	5.03	4.18
Exact (Mult.)	SM03	7.02	9.05
Exact (Mult.)	SM04	6.02	4.97
Exact (Mult.)	SM05	4.59	3.83
Exact (Mult.)	SM06	3.03	1.97
Exact (Mult.)	SM06	11.74	11.20
Exact (Mult.)	SM07	6.08	5.33
Exact (Mult.)	SM08	4.22	5.43
Exact (Mult.)	SM09	5.37	4.16
Exact (Mult.)	SM10	9.02	10.00
Exact (Mult.)	SM11	3.89	3.59
Exact (Mult.)	SM12	5.28	4.28
Exact (Mult.)	SM13	5.77	4.55
Exact (Mult.)	SM14	2.58	1.63
Exact (Mult.)	SM14	5.30	4.84
Exact (Mult.)	SM15	4.70	3.82
Exact (Mult.)	SM15	8.94	9.67
Exact (Mult.)	SM16	5.37	4.88
Exact (Mult.)	SM16	10.65	11.47
Exact (Mult.)	SM17	3.16	3.25
Exact (Mult.)	SM18	2.15	2.58
Exact (Mult.)	SM18	9.58	8.99
Exact (Mult.)	SM18	11.02	10.26
Exact (Mult.)	SM19	9.56	10.39
Exact (Mult.)	SM20	5.70	7.44
Exact (Mult.)	SM21	4.10	3.68
Exact (Mult.)	SM22	2.40	0.94
Exact (Mult.)	SM22	7.43	6.05
Exact (Mult.)	SM23	5.45	5.00
Exact (Mult.)	SM24	2.60	2.09

Table 51.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.18
Exact (NDDO)	SM02	5.03	5.68
Exact (NDDO)	SM03	7.02	7.85
Exact (NDDO)	SM04	6.02	6.66
Exact (NDDO)	SM05	4.59	5.21
Exact (NDDO)	SM06	3.03	3.47
Exact (NDDO)	SM06	11.74	11.17
Exact (NDDO)	SM07	6.08	7.00
Exact (NDDO)	SM08	4.22	5.13
Exact (NDDO)	SM09	5.37	5.60
Exact (NDDO)	SM10	9.02	9.89
Exact (NDDO)	SM11	3.89	4.89
Exact (NDDO)	SM12	5.28	5.64
Exact (NDDO)	SM13	5.77	6.37
Exact (NDDO)	SM14	2.58	2.02
Exact (NDDO)	SM14	5.30	6.38
Exact (NDDO)	SM15	4.70	5.10
Exact (NDDO)	SM15	8.94	10.13
Exact (NDDO)	SM16	5.37	6.55
Exact (NDDO)	SM16	10.65	10.79
Exact (NDDO)	SM17	3.16	5.09
Exact (NDDO)	SM18	2.15	4.58
Exact (NDDO)	SM18	9.58	8.35
Exact (NDDO)	SM18	11.02	11.66
Exact (NDDO)	SM19	9.56	10.05
Exact (NDDO)	SM20	5.70	6.87
Exact (NDDO)	SM21	4.10	5.12
Exact (NDDO)	SM22	2.40	0.21
Exact (NDDO)	SM22	7.43	4.62
Exact (NDDO)	SM23	5.45	6.91
Exact (NDDO)	SM24	2.60	4.09

Table 52.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.71
Exact (NDDO)	SM02	5.03	6.28
Exact (NDDO)	SM03	7.02	7.94
Exact (NDDO)	SM04	6.02	7.03
Exact (NDDO)	SM05	4.59	5.92
Exact (NDDO)	SM06	3.03	4.59
Exact (NDDO)	SM06	11.74	10.48
Exact (NDDO)	SM07	6.08	7.28
Exact (NDDO)	SM08	4.22	5.85
Exact (NDDO)	SM09	5.37	6.21
Exact (NDDO)	SM10	9.02	9.50
Exact (NDDO)	SM11	3.89	5.67
Exact (NDDO)	SM12	5.28	6.25
Exact (NDDO)	SM13	5.77	6.81
Exact (NDDO)	SM14	2.58	3.48
Exact (NDDO)	SM14	5.30	6.81
Exact (NDDO)	SM15	4.70	5.83
Exact (NDDO)	SM15	8.94	9.68
Exact (NDDO)	SM16	5.37	6.94
Exact (NDDO)	SM16	10.65	10.18
Exact (NDDO)	SM17	3.16	5.83
Exact (NDDO)	SM18	2.15	5.44
Exact (NDDO)	SM18	9.58	8.32
Exact (NDDO)	SM18	11.02	10.84
Exact (NDDO)	SM19	9.56	9.61
Exact (NDDO)	SM20	5.70	7.19
Exact (NDDO)	SM21	4.10	5.85
Exact (NDDO)	SM22	2.40	2.10
Exact (NDDO)	SM22	7.43	5.47
Exact (NDDO)	SM23	5.45	7.21
Exact (NDDO)	SM24	2.60	5.06

Table 53.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.54
Exact (NDDO)	SM02	5.03	6.09
Exact (NDDO)	SM03	7.02	7.76
Exact (NDDO)	SM04	6.02	6.84
Exact (NDDO)	SM05	4.59	5.73
Exact (NDDO)	SM06	3.03	4.40
Exact (NDDO)	SM06	11.74	10.30
Exact (NDDO)	SM07	6.08	7.10
Exact (NDDO)	SM08	4.22	5.67
Exact (NDDO)	SM09	5.37	6.03
Exact (NDDO)	SM10	9.02	9.32
Exact (NDDO)	SM11	3.89	5.48
Exact (NDDO)	SM12	5.28	6.06
Exact (NDDO)	SM13	5.77	6.62
Exact (NDDO)	SM14	2.58	3.28
Exact (NDDO)	SM14	5.30	6.63
Exact (NDDO)	SM15	4.70	5.65
Exact (NDDO)	SM15	8.94	9.51
Exact (NDDO)	SM16	5.37	6.76
Exact (NDDO)	SM16	10.65	10.01
Exact (NDDO)	SM17	3.16	5.64
Exact (NDDO)	SM18	2.15	5.25
Exact (NDDO)	SM18	9.58	8.14
Exact (NDDO)	SM18	11.02	10.67
Exact (NDDO)	SM19	9.56	9.44
Exact (NDDO)	SM20	5.70	7.01
Exact (NDDO)	SM21	4.10	5.66
Exact (NDDO)	SM22	2.40	1.90
Exact (NDDO)	SM22	7.43	5.28
Exact (NDDO)	SM23	5.45	7.03
Exact (NDDO)	SM24	2.60	4.87

Table 54.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Point charge	SM01	9.53	6.09	5.08
Point charge	SM02	5.03	3.84	3.07
Point charge	SM03	7.02	6.86	5.97
Point charge	SM04	6.02	4.71	3.99
Point charge	SM05	4.59	3.83	3.28
Point charge	SM06	3.03	3.23	2.56
Point charge	SM06	11.74	10.78	9.86
Point charge	SM07	6.08	4.89	4.14
Point charge	SM08	4.22	3.32	2.31
Point charge	SM09	5.37	3.23	2.36
Point charge	SM10	9.02	8.47	7.65
Point charge	SM11	3.89	3.53	2.87
Point charge	SM12	5.28	3.89	3.06
Point charge	SM13	5.77	3.75	3.00
Point charge	SM14	2.58	1.82	1.24
Point charge	SM14	5.30	4.32	3.65
Point charge	SM15	4.70	4.29	3.54
Point charge	SM15	8.94	6.61	5.66
Point charge	SM16	5.37	5.77	5.10
Point charge	SM16	10.65	10.51	9.64
Point charge	SM17	3.16	4.29	3.67
Point charge	SM18	2.15	2.20	1.54
Point charge	SM18	9.58	7.73	6.86
Point charge	SM18	11.02	8.56	7.55
Point charge	SM19	9.56	9.41	8.60
Point charge	SM20	5.70	6.29	5.45
Point charge	SM21	4.10	5.09	4.36
Point charge	SM22	2.40	2.79	1.95
Point charge	SM22	7.43	5.70	4.92
Point charge	SM23	5.45	3.92	3.22
Point charge	SM24	2.60	2.59	1.91

Table 55.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	5.31
Point charge	SM02	5.03	3.53
Point charge	SM03	7.02	6.09
Point charge	SM04	6.02	4.34
Point charge	SM05	4.59	3.71
Point charge	SM06	3.03	3.08
Point charge	SM06	11.74	9.51
Point charge	SM07	6.08	4.48
Point charge	SM08	4.22	2.87
Point charge	SM09	5.37	2.91
Point charge	SM10	9.02	7.57
Point charge	SM11	3.89	3.36
Point charge	SM12	5.28	3.52
Point charge	SM13	5.77	3.47
Point charge	SM14	2.58	1.92
Point charge	SM14	5.30	4.04
Point charge	SM15	4.70	3.95
Point charge	SM15	8.94	5.81
Point charge	SM16	5.37	5.32
Point charge	SM16	10.65	9.32
Point charge	SM17	3.16	4.06
Point charge	SM18	2.15	2.19
Point charge	SM18	9.58	6.87
Point charge	SM18	11.02	7.48
Point charge	SM19	9.56	8.40
Point charge	SM20	5.70	5.63
Point charge	SM21	4.10	4.67
Point charge	SM22	2.40	2.55
Point charge	SM22	7.43	5.16
Point charge	SM23	5.45	3.67
Point charge	SM24	2.60	2.51

Table 56.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.19.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	4.94
Point charge	SM02	5.03	2.99
Point charge	SM03	7.02	5.79
Point charge	SM04	6.02	3.88
Point charge	SM05	4.59	3.19
Point charge	SM06	3.03	2.50
Point charge	SM06	11.74	9.56
Point charge	SM07	6.08	4.03
Point charge	SM08	4.22	2.26
Point charge	SM09	5.37	2.31
Point charge	SM10	9.02	7.42
Point charge	SM11	3.89	2.80
Point charge	SM12	5.28	2.98
Point charge	SM13	5.77	2.93
Point charge	SM14	2.58	1.22
Point charge	SM14	5.30	3.55
Point charge	SM15	4.70	3.45
Point charge	SM15	8.94	5.49
Point charge	SM16	5.37	4.95
Point charge	SM16	10.65	9.34
Point charge	SM17	3.16	3.57
Point charge	SM18	2.15	1.51
Point charge	SM18	9.58	6.65
Point charge	SM18	11.02	7.33
Point charge	SM19	9.56	8.34
Point charge	SM20	5.70	5.29
Point charge	SM21	4.10	4.24
Point charge	SM22	2.40	1.91
Point charge	SM22	7.43	4.78
Point charge	SM23	5.45	3.14
Point charge	SM24	2.60	1.87

Table 57.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X reoptimised geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.45	10.63
Exact (Mult.)	SM02	5.03	3.38	4.87
Exact (Mult.)	SM03	7.02	9.18	10.48
Exact (Mult.)	SM04	6.02	4.48	5.91
Exact (Mult.)	SM05	4.59	2.46	4.01
Exact (Mult.)	SM06	3.03	1.11	2.62
Exact (Mult.)	SM06	11.74	10.89	12.16
Exact (Mult.)	SM07	6.08	4.20	5.67
Exact (Mult.)	SM08	4.22	5.17	6.31
Exact (Mult.)	SM09	5.37	3.62	5.11
Exact (Mult.)	SM10	9.02	9.49	10.78
Exact (Mult.)	SM11	3.89	2.72	4.27
Exact (Mult.)	SM12	5.28	3.66	5.15
Exact (Mult.)	SM13	5.77	3.35	4.82
Exact (Mult.)	SM14	2.58	0.42	2.04
Exact (Mult.)	SM14	5.30	4.06	5.64
Exact (Mult.)	SM15	4.70	3.28	4.73
Exact (Mult.)	SM15	8.94	9.43	10.65
Exact (Mult.)	SM16	5.37	4.19	5.73
Exact (Mult.)	SM16	10.65	11.48	12.77
Exact (Mult.)	SM17	3.16	2.19	3.80
Exact (Mult.)	SM18	2.15	1.69	3.24
Exact (Mult.)	SM18	9.58	8.76	10.15
Exact (Mult.)	SM18	11.02	10.55	11.62
Exact (Mult.)	SM19	9.56	10.17	11.50
Exact (Mult.)	SM20	5.70	6.78	8.14
Exact (Mult.)	SM21	4.10	3.30	4.79
Exact (Mult.)	SM22	2.40	-0.01	1.59
Exact (Mult.)	SM22	7.43	5.88	7.09
Exact (Mult.)	SM23	5.45	3.62	5.08
Exact (Mult.)	SM24	2.60	1.24	2.73

Table 58.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.20
Exact (Mult.)	SM02	5.03	4.69
Exact (Mult.)	SM03	7.02	9.09
Exact (Mult.)	SM04	6.02	5.50
Exact (Mult.)	SM05	4.59	4.02
Exact (Mult.)	SM06	3.03	2.92
Exact (Mult.)	SM06	11.74	10.41
Exact (Mult.)	SM07	6.08	5.32
Exact (Mult.)	SM08	4.22	5.82
Exact (Mult.)	SM09	5.37	4.88
Exact (Mult.)	SM10	9.02	9.33
Exact (Mult.)	SM11	3.89	4.22
Exact (Mult.)	SM12	5.28	4.91
Exact (Mult.)	SM13	5.77	4.65
Exact (Mult.)	SM14	2.58	2.47
Exact (Mult.)	SM14	5.30	5.29
Exact (Mult.)	SM15	4.70	4.58
Exact (Mult.)	SM15	8.94	9.22
Exact (Mult.)	SM16	5.37	5.36
Exact (Mult.)	SM16	10.65	10.88
Exact (Mult.)	SM17	3.16	3.85
Exact (Mult.)	SM18	2.15	3.41
Exact (Mult.)	SM18	9.58	8.83
Exact (Mult.)	SM18	11.02	9.98
Exact (Mult.)	SM19	9.56	9.89
Exact (Mult.)	SM20	5.70	7.25
Exact (Mult.)	SM21	4.10	4.62
Exact (Mult.)	SM22	2.40	2.11
Exact (Mult.)	SM22	7.43	6.43
Exact (Mult.)	SM23	5.45	4.85
Exact (Mult.)	SM24	2.60	3.01

Table 59.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	9.43
Exact (Mult.)	SM02	5.03	4.10
Exact (Mult.)	SM03	7.02	9.30
Exact (Mult.)	SM04	6.02	5.07
Exact (Mult.)	SM05	4.59	3.31
Exact (Mult.)	SM06	3.03	2.02
Exact (Mult.)	SM06	11.74	10.86
Exact (Mult.)	SM07	6.08	4.85
Exact (Mult.)	SM08	4.22	5.44
Exact (Mult.)	SM09	5.37	4.32
Exact (Mult.)	SM10	9.02	9.58
Exact (Mult.)	SM11	3.89	3.55
Exact (Mult.)	SM12	5.28	4.36
Exact (Mult.)	SM13	5.77	4.05
Exact (Mult.)	SM14	2.58	1.48
Exact (Mult.)	SM14	5.30	4.82
Exact (Mult.)	SM15	4.70	3.98
Exact (Mult.)	SM15	8.94	9.46
Exact (Mult.)	SM16	5.37	4.90
Exact (Mult.)	SM16	10.65	11.42
Exact (Mult.)	SM17	3.16	3.11
Exact (Mult.)	SM18	2.15	2.60
Exact (Mult.)	SM18	9.58	8.99
Exact (Mult.)	SM18	11.02	10.35
Exact (Mult.)	SM19	9.56	10.25
Exact (Mult.)	SM20	5.70	7.13
Exact (Mult.)	SM21	4.10	4.03
Exact (Mult.)	SM22	2.40	1.06
Exact (Mult.)	SM22	7.43	6.15
Exact (Mult.)	SM23	5.45	4.30
Exact (Mult.)	SM24	2.60	2.12

Table 60.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.15
Exact (NDDO)	SM02	5.03	5.71
Exact (NDDO)	SM03	7.02	8.03
Exact (NDDO)	SM04	6.02	6.88
Exact (NDDO)	SM05	4.59	4.78
Exact (NDDO)	SM06	3.03	3.55
Exact (NDDO)	SM06	11.74	10.80
Exact (NDDO)	SM07	6.08	6.46
Exact (NDDO)	SM08	4.22	5.13
Exact (NDDO)	SM09	5.37	5.84
Exact (NDDO)	SM10	9.02	9.46
Exact (NDDO)	SM11	3.89	4.86
Exact (NDDO)	SM12	5.28	5.80
Exact (NDDO)	SM13	5.77	5.90
Exact (NDDO)	SM14	2.58	1.86
Exact (NDDO)	SM14	5.30	6.36
Exact (NDDO)	SM15	4.70	5.26
Exact (NDDO)	SM15	8.94	9.91
Exact (NDDO)	SM16	5.37	6.57
Exact (NDDO)	SM16	10.65	10.68
Exact (NDDO)	SM17	3.16	5.01
Exact (NDDO)	SM18	2.15	4.57
Exact (NDDO)	SM18	9.58	8.35
Exact (NDDO)	SM18	11.02	11.79
Exact (NDDO)	SM19	9.56	9.91
Exact (NDDO)	SM20	5.70	6.49
Exact (NDDO)	SM21	4.10	5.46
Exact (NDDO)	SM22	2.40	0.40
Exact (NDDO)	SM22	7.43	4.70
Exact (NDDO)	SM23	5.45	6.17
Exact (NDDO)	SM24	2.60	4.17

Table 61.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.69
Exact (NDDO)	SM02	5.03	6.30
Exact (NDDO)	SM03	7.02	8.07
Exact (NDDO)	SM04	6.02	7.19
Exact (NDDO)	SM05	4.59	5.59
Exact (NDDO)	SM06	3.03	4.65
Exact (NDDO)	SM06	11.74	10.19
Exact (NDDO)	SM07	6.08	6.88
Exact (NDDO)	SM08	4.22	5.85
Exact (NDDO)	SM09	5.37	6.40
Exact (NDDO)	SM10	9.02	9.17
Exact (NDDO)	SM11	3.89	5.65
Exact (NDDO)	SM12	5.28	6.37
Exact (NDDO)	SM13	5.77	6.45
Exact (NDDO)	SM14	2.58	3.36
Exact (NDDO)	SM14	5.30	6.80
Exact (NDDO)	SM15	4.70	5.96
Exact (NDDO)	SM15	8.94	9.51
Exact (NDDO)	SM16	5.37	6.96
Exact (NDDO)	SM16	10.65	10.10
Exact (NDDO)	SM17	3.16	5.77
Exact (NDDO)	SM18	2.15	5.43
Exact (NDDO)	SM18	9.58	8.32
Exact (NDDO)	SM18	11.02	10.95
Exact (NDDO)	SM19	9.56	9.51
Exact (NDDO)	SM20	5.70	6.90
Exact (NDDO)	SM21	4.10	6.11
Exact (NDDO)	SM22	2.40	2.25
Exact (NDDO)	SM22	7.43	5.53
Exact (NDDO)	SM23	5.45	6.65
Exact (NDDO)	SM24	2.60	5.12

Table 62.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.52
Exact (NDDO)	SM02	5.03	6.11
Exact (NDDO)	SM03	7.02	7.89
Exact (NDDO)	SM04	6.02	7.01
Exact (NDDO)	SM05	4.59	5.40
Exact (NDDO)	SM06	3.03	4.46
Exact (NDDO)	SM06	11.74	10.02
Exact (NDDO)	SM07	6.08	6.69
Exact (NDDO)	SM08	4.22	5.67
Exact (NDDO)	SM09	5.37	6.21
Exact (NDDO)	SM10	9.02	8.99
Exact (NDDO)	SM11	3.89	5.47
Exact (NDDO)	SM12	5.28	6.18
Exact (NDDO)	SM13	5.77	6.26
Exact (NDDO)	SM14	2.58	3.17
Exact (NDDO)	SM14	5.30	6.61
Exact (NDDO)	SM15	4.70	5.77
Exact (NDDO)	SM15	8.94	9.34
Exact (NDDO)	SM16	5.37	6.77
Exact (NDDO)	SM16	10.65	9.92
Exact (NDDO)	SM17	3.16	5.58
Exact (NDDO)	SM18	2.15	5.24
Exact (NDDO)	SM18	9.58	8.14
Exact (NDDO)	SM18	11.02	10.77
Exact (NDDO)	SM19	9.56	9.34
Exact (NDDO)	SM20	5.70	6.71
Exact (NDDO)	SM21	4.10	5.92
Exact (NDDO)	SM22	2.40	2.05
Exact (NDDO)	SM22	7.43	5.34
Exact (NDDO)	SM23	5.45	6.47
Exact (NDDO)	SM24	2.60	4.94

Table 63.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Point charge	SM01	9.53	6.08	4.91
Point charge	SM02	5.03	4.05	3.14
Point charge	SM03	7.02	7.00	5.94
Point charge	SM04	6.02	4.96	4.03
Point charge	SM05	4.59	3.44	2.60
Point charge	SM06	3.03	3.22	2.37
Point charge	SM06	11.74	10.02	8.96
Point charge	SM07	6.08	4.42	3.52
Point charge	SM08	4.22	3.38	2.21
Point charge	SM09	5.37	3.78	2.76
Point charge	SM10	9.02	7.93	6.91
Point charge	SM11	3.89	3.49	2.67
Point charge	SM12	5.28	3.95	2.95
Point charge	SM13	5.77	3.72	2.79
Point charge	SM14	2.58	1.70	0.93
Point charge	SM14	5.30	4.33	3.49
Point charge	SM15	4.70	4.42	3.52
Point charge	SM15	8.94	6.30	5.18
Point charge	SM16	5.37	5.83	5.00
Point charge	SM16	10.65	10.26	9.23
Point charge	SM17	3.16	4.23	3.46
Point charge	SM18	2.15	2.26	1.42
Point charge	SM18	9.58	7.76	6.76
Point charge	SM18	11.02	8.53	7.37
Point charge	SM19	9.56	9.20	8.22
Point charge	SM20	5.70	5.77	4.77
Point charge	SM21	4.10	5.24	4.32
Point charge	SM22	2.40	2.92	1.95
Point charge	SM22	7.43	5.68	4.72
Point charge	SM23	5.45	3.32	2.40
Point charge	SM24	2.60	2.67	1.82

Table 64.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Point charge	SM01	9.53	5.29
Point charge	SM02	5.03	3.73
Point charge	SM03	7.02	6.20
Point charge	SM04	6.02	4.52
Point charge	SM05	4.59	3.26
Point charge	SM06	3.03	3.06
Point charge	SM06	11.74	8.86
Point charge	SM07	6.08	4.07
Point charge	SM08	4.22	2.91
Point charge	SM09	5.37	3.39
Point charge	SM10	9.02	7.06
Point charge	SM11	3.89	3.32
Point charge	SM12	5.28	3.57
Point charge	SM13	5.77	3.43
Point charge	SM14	2.58	1.79
Point charge	SM14	5.30	4.05
Point charge	SM15	4.70	4.07
Point charge	SM15	8.94	5.53
Point charge	SM16	5.37	5.37
Point charge	SM16	10.65	9.10
Point charge	SM17	3.16	4.02
Point charge	SM18	2.15	2.23
Point charge	SM18	9.58	6.93
Point charge	SM18	11.02	7.45
Point charge	SM19	9.56	8.21
Point charge	SM20	5.70	5.17
Point charge	SM21	4.10	4.78
Point charge	SM22	2.40	2.68
Point charge	SM22	7.43	5.13
Point charge	SM23	5.45	3.09
Point charge	SM24	2.60	2.58

Table 65.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.20.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Point charge	SM01	9.53	4.92
Point charge	SM02	5.03	3.21
Point charge	SM03	7.02	5.92
Point charge	SM04	6.02	4.07
Point charge	SM05	4.59	2.69
Point charge	SM06	3.03	2.48
Point charge	SM06	11.74	8.84
Point charge	SM07	6.08	3.58
Point charge	SM08	4.22	2.31
Point charge	SM09	5.37	2.84
Point charge	SM10	9.02	6.86
Point charge	SM11	3.89	2.76
Point charge	SM12	5.28	3.03
Point charge	SM13	5.77	2.88
Point charge	SM14	2.58	1.08
Point charge	SM14	5.30	3.56
Point charge	SM15	4.70	3.58
Point charge	SM15	8.94	5.18
Point charge	SM16	5.37	5.01
Point charge	SM16	10.65	9.10
Point charge	SM17	3.16	3.53
Point charge	SM18	2.15	1.56
Point charge	SM18	9.58	6.72
Point charge	SM18	11.02	7.30
Point charge	SM19	9.56	8.12
Point charge	SM20	5.70	4.79
Point charge	SM21	4.10	4.36
Point charge	SM22	2.40	2.06
Point charge	SM22	7.43	4.75
Point charge	SM23	5.45	2.50
Point charge	SM24	2.60	1.94

Table 66.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Exact (Mult.)	SM01	9.53	10.30	11.51
Exact (Mult.)	SM02	5.03	2.40	3.87
Exact (Mult.)	SM03	7.02	14.09	15.19
Exact (Mult.)	SM04	6.02	3.41	4.82
Exact (Mult.)	SM05	4.59	2.88	4.51
Exact (Mult.)	SM06	3.03	0.99	2.47
Exact (Mult.)	SM06	11.74	10.05	11.32
Exact (Mult.)	SM07	6.08	3.19	4.65
Exact (Mult.)	SM08	4.22	3.45	4.63
Exact (Mult.)	SM09	5.37	2.71	4.17
Exact (Mult.)	SM10	9.02	11.65	13.11
Exact (Mult.)	SM11	3.89	1.41	2.96
Exact (Mult.)	SM12	5.28	2.70	4.17
Exact (Mult.)	SM13	5.77	2.51	3.97
Exact (Mult.)	SM14	2.58	-0.06	1.48
Exact (Mult.)	SM14	5.30	3.87	5.56
Exact (Mult.)	SM15	4.70	2.70	4.12
Exact (Mult.)	SM15	8.94	10.13	11.38
Exact (Mult.)	SM16	5.37	3.35	4.84
Exact (Mult.)	SM16	10.65	10.68	11.95
Exact (Mult.)	SM17	3.16	1.38	2.98
Exact (Mult.)	SM18	2.15	3.61	4.94
Exact (Mult.)	SM18	9.58	7.70	9.08
Exact (Mult.)	SM18	11.02	12.53	13.54
Exact (Mult.)	SM19	9.56	13.19	14.36
Exact (Mult.)	SM20	5.70	7.88	9.23
Exact (Mult.)	SM21	4.10	2.36	3.80
Exact (Mult.)	SM22	2.40	-0.69	0.85
Exact (Mult.)	SM22	7.43	5.58	6.84
Exact (Mult.)	SM23	5.45	2.55	4.01
Exact (Mult.)	SM24	2.60	1.07	2.48

Table 67.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Exact (Mult.)	SM01	9.53	10.30
Exact (Mult.)	SM02	5.03	4.49
Exact (Mult.)	SM03	7.02	13.10
Exact (Mult.)	SM04	6.02	5.21
Exact (Mult.)	SM05	4.59	4.97
Exact (Mult.)	SM06	3.03	3.42
Exact (Mult.)	SM06	11.74	10.16
Exact (Mult.)	SM07	6.08	5.08
Exact (Mult.)	SM08	4.22	5.07
Exact (Mult.)	SM09	5.37	4.72
Exact (Mult.)	SM10	9.02	11.52
Exact (Mult.)	SM11	3.89	3.79
Exact (Mult.)	SM12	5.28	4.72
Exact (Mult.)	SM13	5.77	4.56
Exact (Mult.)	SM14	2.58	2.67
Exact (Mult.)	SM14	5.30	5.77
Exact (Mult.)	SM15	4.70	4.67
Exact (Mult.)	SM15	8.94	10.20
Exact (Mult.)	SM16	5.37	5.23
Exact (Mult.)	SM16	10.65	10.63
Exact (Mult.)	SM17	3.16	3.81
Exact (Mult.)	SM18	2.15	5.30
Exact (Mult.)	SM18	9.58	8.45
Exact (Mult.)	SM18	11.02	11.84
Exact (Mult.)	SM19	9.56	12.47
Exact (Mult.)	SM20	5.70	8.57
Exact (Mult.)	SM21	4.10	4.43
Exact (Mult.)	SM22	2.40	2.19
Exact (Mult.)	SM22	7.43	6.75
Exact (Mult.)	SM23	5.45	4.60
Exact (Mult.)	SM24	2.60	3.43

Table 68.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Exact (Mult.)	SM01	9.53	10.52
Exact (Mult.)	SM02	5.03	3.97
Exact (Mult.)	SM03	7.02	13.69
Exact (Mult.)	SM04	6.02	4.79
Exact (Mult.)	SM05	4.59	4.52
Exact (Mult.)	SM06	3.03	2.77
Exact (Mult.)	SM06	11.74	10.37
Exact (Mult.)	SM07	6.08	4.64
Exact (Mult.)	SM08	4.22	4.62
Exact (Mult.)	SM09	5.37	4.23
Exact (Mult.)	SM10	9.02	11.90
Exact (Mult.)	SM11	3.89	3.19
Exact (Mult.)	SM12	5.28	4.23
Exact (Mult.)	SM13	5.77	4.05
Exact (Mult.)	SM14	2.58	1.92
Exact (Mult.)	SM14	5.30	5.42
Exact (Mult.)	SM15	4.70	4.18
Exact (Mult.)	SM15	8.94	10.42
Exact (Mult.)	SM16	5.37	4.81
Exact (Mult.)	SM16	10.65	10.90
Exact (Mult.)	SM17	3.16	3.20
Exact (Mult.)	SM18	2.15	4.89
Exact (Mult.)	SM18	9.58	8.44
Exact (Mult.)	SM18	11.02	12.27
Exact (Mult.)	SM19	9.56	12.97
Exact (Mult.)	SM20	5.70	8.57
Exact (Mult.)	SM21	4.10	3.91
Exact (Mult.)	SM22	2.40	1.38
Exact (Mult.)	SM22	7.43	6.52
Exact (Mult.)	SM23	5.45	4.09
Exact (Mult.)	SM24	2.60	2.78

Table 69.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.33
Exact (NDDO)	SM02	5.03	4.51
Exact (NDDO)	SM03	7.02	12.86
Exact (NDDO)	SM04	6.02	5.60
Exact (NDDO)	SM05	4.59	4.83
Exact (NDDO)	SM06	3.03	3.21
Exact (NDDO)	SM06	11.74	9.19
Exact (NDDO)	SM07	6.08	5.27
Exact (NDDO)	SM08	4.22	2.72
Exact (NDDO)	SM09	5.37	4.69
Exact (NDDO)	SM10	9.02	11.10
Exact (NDDO)	SM11	3.89	3.37
Exact (NDDO)	SM12	5.28	4.61
Exact (NDDO)	SM13	5.77	4.94
Exact (NDDO)	SM14	2.58	0.78
Exact (NDDO)	SM14	5.30	6.03
Exact (NDDO)	SM15	4.70	4.39
Exact (NDDO)	SM15	8.94	9.97
Exact (NDDO)	SM16	5.37	5.43
Exact (NDDO)	SM16	10.65	9.18
Exact (NDDO)	SM17	3.16	3.90
Exact (NDDO)	SM18	2.15	6.31
Exact (NDDO)	SM18	9.58	6.58
Exact (NDDO)	SM18	11.02	12.92
Exact (NDDO)	SM19	9.56	12.06
Exact (NDDO)	SM20	5.70	6.77
Exact (NDDO)	SM21	4.10	4.37
Exact (NDDO)	SM22	2.40	-0.54
Exact (NDDO)	SM22	7.43	3.60
Exact (NDDO)	SM23	5.45	5.02
Exact (NDDO)	SM24	2.60	3.86

Table 70.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	10.10
Exact (NDDO)	SM02	5.03	5.73
Exact (NDDO)	SM03	7.02	11.99
Exact (NDDO)	SM04	6.02	6.54
Exact (NDDO)	SM05	4.59	5.96
Exact (NDDO)	SM06	3.03	4.75
Exact (NDDO)	SM06	11.74	9.24
Exact (NDDO)	SM07	6.08	6.30
Exact (NDDO)	SM08	4.22	4.38
Exact (NDDO)	SM09	5.37	5.87
Exact (NDDO)	SM10	9.02	10.67
Exact (NDDO)	SM11	3.89	4.87
Exact (NDDO)	SM12	5.28	5.81
Exact (NDDO)	SM13	5.77	6.05
Exact (NDDO)	SM14	2.58	2.92
Exact (NDDO)	SM14	5.30	6.86
Exact (NDDO)	SM15	4.70	5.64
Exact (NDDO)	SM15	8.94	9.82
Exact (NDDO)	SM16	5.37	6.41
Exact (NDDO)	SM16	10.65	9.23
Exact (NDDO)	SM17	3.16	5.27
Exact (NDDO)	SM18	2.15	7.08
Exact (NDDO)	SM18	9.58	7.28
Exact (NDDO)	SM18	11.02	12.04
Exact (NDDO)	SM19	9.56	11.40
Exact (NDDO)	SM20	5.70	7.43
Exact (NDDO)	SM21	4.10	5.62
Exact (NDDO)	SM22	2.40	1.94
Exact (NDDO)	SM22	7.43	5.05
Exact (NDDO)	SM23	5.45	6.11
Exact (NDDO)	SM24	2.60	5.24

Table 71.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (Indiv.,+)
Exact (NDDO)	SM01	9.53	9.90
Exact (NDDO)	SM02	5.03	5.57
Exact (NDDO)	SM03	7.02	11.79
Exact (NDDO)	SM04	6.02	6.38
Exact (NDDO)	SM05	4.59	5.80
Exact (NDDO)	SM06	3.03	4.60
Exact (NDDO)	SM06	11.74	9.06
Exact (NDDO)	SM07	6.08	6.13
Exact (NDDO)	SM08	4.22	4.23
Exact (NDDO)	SM09	5.37	5.71
Exact (NDDO)	SM10	9.02	10.48
Exact (NDDO)	SM11	3.89	4.72
Exact (NDDO)	SM12	5.28	5.65
Exact (NDDO)	SM13	5.77	5.89
Exact (NDDO)	SM14	2.58	2.79
Exact (NDDO)	SM14	5.30	6.70
Exact (NDDO)	SM15	4.70	5.48
Exact (NDDO)	SM15	8.94	9.63
Exact (NDDO)	SM16	5.37	6.25
Exact (NDDO)	SM16	10.65	9.04
Exact (NDDO)	SM17	3.16	5.11
Exact (NDDO)	SM18	2.15	6.91
Exact (NDDO)	SM18	9.58	7.11
Exact (NDDO)	SM18	11.02	11.83
Exact (NDDO)	SM19	9.56	11.20
Exact (NDDO)	SM20	5.70	7.25
Exact (NDDO)	SM21	4.10	5.47
Exact (NDDO)	SM22	2.40	1.81
Exact (NDDO)	SM22	7.43	4.89
Exact (NDDO)	SM23	5.45	5.95
Exact (NDDO)	SM24	2.60	5.08

Table 72.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$	
			hr-MP2 (All)	/B@PFL (All)
Point charge	SM01	9.53	6.87	5.72
Point charge	SM02	5.03	3.21	2.27
Point charge	SM03	7.02	11.82	10.73
Point charge	SM04	6.02	4.10	3.14
Point charge	SM05	4.59	3.80	3.02
Point charge	SM06	3.03	3.20	2.35
Point charge	SM06	11.74	8.91	7.81
Point charge	SM07	6.08	3.63	2.71
Point charge	SM08	4.22	1.66	0.54
Point charge	SM09	5.37	2.95	1.93
Point charge	SM10	9.02	9.73	8.82
Point charge	SM11	3.89	2.40	1.56
Point charge	SM12	5.28	3.22	2.25
Point charge	SM13	5.77	3.02	2.11
Point charge	SM14	2.58	1.31	0.47
Point charge	SM14	5.30	4.08	3.33
Point charge	SM15	4.70	3.84	2.91
Point charge	SM15	8.94	7.00	5.88
Point charge	SM16	5.37	4.99	4.12
Point charge	SM16	10.65	9.52	8.47
Point charge	SM17	3.16	3.43	2.62
Point charge	SM18	2.15	4.55	3.61
Point charge	SM18	9.58	7.00	5.99
Point charge	SM18	11.02	10.83	9.63
Point charge	SM19	9.56	12.19	11.11
Point charge	SM20	5.70	6.72	5.73
Point charge	SM21	4.10	4.34	3.38
Point charge	SM22	2.40	2.73	1.88
Point charge	SM22	7.43	4.93	3.84
Point charge	SM23	5.45	2.56	1.66
Point charge	SM24	2.60	2.29	1.37

Table 73.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$ /B@PFL (All)
Point charge	SM01	9.53	6.51
Point charge	SM02	5.03	3.53
Point charge	SM03	7.02	10.84
Point charge	SM04	6.02	4.27
Point charge	SM05	4.59	4.17
Point charge	SM06	3.03	3.59
Point charge	SM06	11.74	8.31
Point charge	SM07	6.08	3.91
Point charge	SM08	4.22	2.03
Point charge	SM09	5.37	3.23
Point charge	SM10	9.02	9.19
Point charge	SM11	3.89	2.91
Point charge	SM12	5.28	3.50
Point charge	SM13	5.77	3.39
Point charge	SM14	2.58	1.97
Point charge	SM14	5.30	4.44
Point charge	SM15	4.70	4.07
Point charge	SM15	8.94	6.65
Point charge	SM16	5.37	5.13
Point charge	SM16	10.65	8.88
Point charge	SM17	3.16	3.83
Point charge	SM18	2.15	4.68
Point charge	SM18	9.58	6.74
Point charge	SM18	11.02	9.88
Point charge	SM19	9.56	11.16
Point charge	SM20	5.70	6.51
Point charge	SM21	4.10	4.49
Point charge	SM22	2.40	3.19
Point charge	SM22	7.43	4.88
Point charge	SM23	5.45	3.00
Point charge	SM24	2.60	2.74

Table 74.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.21.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$
			/B@PFL (All)
Point charge	SM01	9.53	6.26
Point charge	SM02	5.03	3.08
Point charge	SM03	7.02	10.88
Point charge	SM04	6.02	3.88
Point charge	SM05	4.59	3.77
Point charge	SM06	3.03	3.15
Point charge	SM06	11.74	8.19
Point charge	SM07	6.08	3.49
Point charge	SM08	4.22	1.48
Point charge	SM09	5.37	2.77
Point charge	SM10	9.02	9.12
Point charge	SM11	3.89	2.43
Point charge	SM12	5.28	3.06
Point charge	SM13	5.77	2.94
Point charge	SM14	2.58	1.42
Point charge	SM14	5.30	4.06
Point charge	SM15	4.70	3.67
Point charge	SM15	8.94	6.41
Point charge	SM16	5.37	4.79
Point charge	SM16	10.65	8.79
Point charge	SM17	3.16	3.41
Point charge	SM18	2.15	4.32
Point charge	SM18	9.58	6.51
Point charge	SM18	11.02	9.86
Point charge	SM19	9.56	11.22
Point charge	SM20	5.70	6.27
Point charge	SM21	4.10	4.11
Point charge	SM22	2.40	2.72
Point charge	SM22	7.43	4.53
Point charge	SM23	5.45	2.52
Point charge	SM24	2.60	2.25

Table 75.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.83	11.13	9.82	9.34
Exact (Mult.)	SM02	5.03	3.88	5.48	3.82	5.18
Exact (Mult.)	SM03	7.02	8.52	9.94	8.52	6.87
Exact (Mult.)	SM04	6.02	4.57	6.18	4.57	5.88
Exact (Mult.)	SM05	4.59	3.39	5.24	3.44	4.63
Exact (Mult.)	SM06	3.03	1.15	2.78	1.16	1.34
Exact (Mult.)	SM06	11.74	11.53	12.90	11.55	11.36
Exact (Mult.)	SM07	6.08	5.15	6.73	5.10	6.26
Exact (Mult.)	SM08	4.22	5.20	6.48	5.15	4.04
Exact (Mult.)	SM09	5.37	3.56	5.16	3.52	4.72
Exact (Mult.)	SM10	9.02	9.88	11.34	9.91	9.00
Exact (Mult.)	SM11	3.89	3.03	4.70	3.00	3.95
Exact (Mult.)	SM12	5.28	3.70	5.30	3.66	4.75
Exact (Mult.)	SM13	5.77	4.14	5.73	4.08	5.78
Exact (Mult.)	SM14	2.58	0.68	2.48	0.60	0.62
Exact (Mult.)	SM14	5.30	4.21	5.90	4.15	5.10
Exact (Mult.)	SM15	4.70	3.36	4.92	3.31	3.84
Exact (Mult.)	SM15	8.94	9.67	11.04	9.69	9.17
Exact (Mult.)	SM16	5.37	4.24	5.91	4.26	5.50
Exact (Mult.)	SM16	10.65	11.61	13.01	11.64	10.38
Exact (Mult.)	SM17	3.16	2.42	4.14	2.42	3.60
Exact (Mult.)	SM18	2.15	1.96	3.62	1.85	3.67
Exact (Mult.)	SM18	9.58	9.14	10.65	9.09	7.32
Exact (Mult.)	SM18	11.02	10.65	11.85	10.67	10.66
Exact (Mult.)	SM19	9.56	10.65	12.11	10.72	9.00
Exact (Mult.)	SM20	5.70	7.10	8.54	7.10	4.79
Exact (Mult.)	SM21	4.10	3.33	4.93	3.37	3.71
Exact (Mult.)	SM22	2.40	0.62	2.40	0.65	1.00
Exact (Mult.)	SM22	7.43	7.14	8.45	7.24	6.85
Exact (Mult.)	SM23	5.45	5.12	6.76	5.15	8.38
Exact (Mult.)	SM24	2.60	1.38	3.02	1.33	3.92

Table 76.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.51	8.48	8.11
Exact (Mult.)	SM02	5.03	5.08	3.77	4.84
Exact (Mult.)	SM03	7.02	8.58	7.46	6.16
Exact (Mult.)	SM04	6.02	5.62	4.36	5.39
Exact (Mult.)	SM05	4.59	4.88	3.47	4.41
Exact (Mult.)	SM06	3.03	2.95	1.69	1.83
Exact (Mult.)	SM06	11.74	10.90	9.84	9.69
Exact (Mult.)	SM07	6.08	6.05	4.78	5.69
Exact (Mult.)	SM08	4.22	5.86	4.82	3.95
Exact (Mult.)	SM09	5.37	4.83	3.53	4.48
Exact (Mult.)	SM10	9.02	9.67	8.55	7.83
Exact (Mult.)	SM11	3.89	4.47	3.13	3.87
Exact (Mult.)	SM12	5.28	4.93	3.65	4.50
Exact (Mult.)	SM13	5.77	5.27	3.97	5.31
Exact (Mult.)	SM14	2.58	2.72	1.24	1.26
Exact (Mult.)	SM14	5.30	5.40	4.03	4.78
Exact (Mult.)	SM15	4.70	4.64	3.37	3.79
Exact (Mult.)	SM15	8.94	9.43	8.38	7.97
Exact (Mult.)	SM16	5.37	5.41	4.12	5.09
Exact (Mult.)	SM16	10.65	10.98	9.90	8.92
Exact (Mult.)	SM17	3.16	4.02	2.67	3.60
Exact (Mult.)	SM18	2.15	3.62	2.22	3.65
Exact (Mult.)	SM18	9.58	9.13	7.91	6.52
Exact (Mult.)	SM18	11.02	10.08	9.15	9.14
Exact (Mult.)	SM19	9.56	10.27	9.18	7.84
Exact (Mult.)	SM20	5.70	7.47	6.34	4.54
Exact (Mult.)	SM21	4.10	4.64	3.42	3.69
Exact (Mult.)	SM22	2.40	2.66	1.29	1.56
Exact (Mult.)	SM22	7.43	7.41	6.45	6.15
Exact (Mult.)	SM23	5.45	6.08	4.82	7.35
Exact (Mult.)	SM24	2.60	3.14	1.81	3.85

Table 77.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.80	8.58	8.14
Exact (Mult.)	SM02	5.03	4.56	3.02	4.28
Exact (Mult.)	SM03	7.02	8.69	7.37	5.84
Exact (Mult.)	SM04	6.02	5.21	3.71	4.93
Exact (Mult.)	SM05	4.59	4.33	2.67	3.77
Exact (Mult.)	SM06	3.03	2.05	0.56	0.72
Exact (Mult.)	SM06	11.74	11.43	10.19	10.01
Exact (Mult.)	SM07	6.08	5.71	4.21	5.28
Exact (Mult.)	SM08	4.22	5.49	4.26	3.23
Exact (Mult.)	SM09	5.37	4.26	2.74	3.86
Exact (Mult.)	SM10	9.02	9.99	8.66	7.82
Exact (Mult.)	SM11	3.89	3.84	2.26	3.14
Exact (Mult.)	SM12	5.28	4.39	2.87	3.88
Exact (Mult.)	SM13	5.77	4.79	3.26	4.84
Exact (Mult.)	SM14	2.58	1.78	0.03	0.06
Exact (Mult.)	SM14	5.30	4.95	3.33	4.21
Exact (Mult.)	SM15	4.70	4.04	2.55	3.04
Exact (Mult.)	SM15	8.94	9.71	8.46	7.98
Exact (Mult.)	SM16	5.37	4.95	3.43	4.58
Exact (Mult.)	SM16	10.65	11.53	10.26	9.10
Exact (Mult.)	SM17	3.16	3.31	1.72	2.82
Exact (Mult.)	SM18	2.15	2.84	1.19	2.88
Exact (Mult.)	SM18	9.58	9.35	7.91	6.27
Exact (Mult.)	SM18	11.02	10.47	9.37	9.36
Exact (Mult.)	SM19	9.56	10.70	9.41	7.82
Exact (Mult.)	SM20	5.70	7.39	6.06	3.92
Exact (Mult.)	SM21	4.10	4.05	2.60	2.92
Exact (Mult.)	SM22	2.40	1.71	0.09	0.41
Exact (Mult.)	SM22	7.43	7.31	6.18	5.82
Exact (Mult.)	SM23	5.45	5.74	4.25	7.25
Exact (Mult.)	SM24	2.60	2.28	0.71	3.11

Table 78.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	10.33	9.91	9.57
Exact (NDDO)	SM02	5.03	6.02	3.39	4.37
Exact (NDDO)	SM03	7.02	7.75	7.58	6.41
Exact (NDDO)	SM04	6.02	6.98	4.41	5.34
Exact (NDDO)	SM05	4.59	6.51	3.58	4.42
Exact (NDDO)	SM06	3.03	3.54	0.98	1.07
Exact (NDDO)	SM06	11.74	11.41	11.25	11.15
Exact (NDDO)	SM07	6.08	7.47	4.95	5.78
Exact (NDDO)	SM08	4.22	4.99	4.95	4.15
Exact (NDDO)	SM09	5.37	5.68	3.13	3.99
Exact (NDDO)	SM10	9.02	9.90	9.76	9.12
Exact (NDDO)	SM11	3.89	5.16	2.81	3.48
Exact (NDDO)	SM12	5.28	5.90	3.40	4.16
Exact (NDDO)	SM13	5.77	6.58	3.82	5.04
Exact (NDDO)	SM14	2.58	2.33	0.60	0.62
Exact (NDDO)	SM14	5.30	6.56	4.08	4.76
Exact (NDDO)	SM15	4.70	5.46	3.12	3.49
Exact (NDDO)	SM15	8.94	10.11	9.73	9.35
Exact (NDDO)	SM16	5.37	6.58	4.09	4.97
Exact (NDDO)	SM16	10.65	11.05	10.75	9.87
Exact (NDDO)	SM17	3.16	5.08	2.40	3.24
Exact (NDDO)	SM18	2.15	4.82	1.59	2.90
Exact (NDDO)	SM18	9.58	8.54	8.69	7.41
Exact (NDDO)	SM18	11.02	11.70	10.60	10.60
Exact (NDDO)	SM19	9.56	10.16	10.28	9.07
Exact (NDDO)	SM20	5.70	6.87	6.85	5.22
Exact (NDDO)	SM21	4.10	5.77	3.07	3.26
Exact (NDDO)	SM22	2.40	1.26	-0.94	-0.75
Exact (NDDO)	SM22	7.43	5.47	4.94	4.72
Exact (NDDO)	SM23	5.45	7.45	4.21	6.54
Exact (NDDO)	SM24	2.60	4.38	1.24	3.10

Table 79.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			/B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	8.57	9.83	9.51	9.25
Exact (NDDO)	SM02	5.03	3.84	6.54	4.53	5.28
Exact (NDDO)	SM03	7.02	6.87	7.86	7.73	6.84
Exact (NDDO)	SM04	6.02	4.59	7.27	5.31	6.02
Exact (NDDO)	SM05	4.59	4.09	6.91	4.67	5.32
Exact (NDDO)	SM06	3.03	1.96	4.64	2.69	2.76
Exact (NDDO)	SM06	11.74	9.62	10.66	10.53	10.45
Exact (NDDO)	SM07	6.08	5.01	7.64	5.72	6.35
Exact (NDDO)	SM08	4.22	4.80	5.75	5.72	5.11
Exact (NDDO)	SM09	5.37	3.64	6.28	4.33	4.99
Exact (NDDO)	SM10	9.02	8.55	9.50	9.40	8.90
Exact (NDDO)	SM11	3.89	3.44	5.88	4.08	4.60
Exact (NDDO)	SM12	5.28	3.83	6.44	4.53	5.12
Exact (NDDO)	SM13	5.77	4.17	6.96	4.86	5.79
Exact (NDDO)	SM14	2.58	1.88	3.72	2.40	2.41
Exact (NDDO)	SM14	5.30	4.44	6.95	5.06	5.57
Exact (NDDO)	SM15	4.70	3.60	6.11	4.32	4.61
Exact (NDDO)	SM15	8.94	8.45	9.66	9.37	9.08
Exact (NDDO)	SM16	5.37	4.38	6.96	5.07	5.73
Exact (NDDO)	SM16	10.65	9.26	10.38	10.15	9.48
Exact (NDDO)	SM17	3.16	3.14	5.82	3.77	4.42
Exact (NDDO)	SM18	2.15	2.55	5.62	3.15	4.15
Exact (NDDO)	SM18	9.58	7.82	8.46	8.58	7.60
Exact (NDDO)	SM18	11.02	9.00	10.88	10.04	10.03
Exact (NDDO)	SM19	9.56	8.92	9.70	9.79	8.87
Exact (NDDO)	SM20	5.70	6.33	7.18	7.17	5.93
Exact (NDDO)	SM21	4.10	3.53	6.35	4.29	4.43
Exact (NDDO)	SM22	2.40	0.50	2.90	1.22	1.36
Exact (NDDO)	SM22	7.43	4.80	6.12	5.71	5.55
Exact (NDDO)	SM23	5.45	4.44	7.63	5.15	6.94
Exact (NDDO)	SM24	2.60	2.23	5.29	2.89	4.31

Table 80.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			/B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	8.39	9.65	9.33	9.07
Exact (NDDO)	SM02	5.03	3.65	6.35	4.34	5.09
Exact (NDDO)	SM03	7.02	6.69	7.68	7.55	6.65
Exact (NDDO)	SM04	6.02	4.40	7.08	5.12	5.83
Exact (NDDO)	SM05	4.59	3.89	6.73	4.48	5.13
Exact (NDDO)	SM06	3.03	1.76	4.45	2.49	2.56
Exact (NDDO)	SM06	11.74	9.45	10.49	10.36	10.28
Exact (NDDO)	SM07	6.08	4.83	7.46	5.54	6.17
Exact (NDDO)	SM08	4.22	4.61	5.56	5.53	4.92
Exact (NDDO)	SM09	5.37	3.44	6.09	4.14	4.80
Exact (NDDO)	SM10	9.02	8.37	9.33	9.22	8.73
Exact (NDDO)	SM11	3.89	3.24	5.70	3.89	4.41
Exact (NDDO)	SM12	5.28	3.64	6.26	4.34	4.93
Exact (NDDO)	SM13	5.77	3.97	6.78	4.67	5.61
Exact (NDDO)	SM14	2.58	1.68	3.52	2.20	2.21
Exact (NDDO)	SM14	5.30	4.25	6.77	4.87	5.39
Exact (NDDO)	SM15	4.70	3.41	5.93	4.13	4.42
Exact (NDDO)	SM15	8.94	8.27	9.49	9.19	8.91
Exact (NDDO)	SM16	5.37	4.18	6.78	4.88	5.54
Exact (NDDO)	SM16	10.65	9.08	10.21	9.98	9.31
Exact (NDDO)	SM17	3.16	2.94	5.63	3.58	4.22
Exact (NDDO)	SM18	2.15	2.36	5.43	2.96	3.96
Exact (NDDO)	SM18	9.58	7.64	8.28	8.40	7.42
Exact (NDDO)	SM18	11.02	8.82	10.70	9.86	9.86
Exact (NDDO)	SM19	9.56	8.74	9.52	9.62	8.69
Exact (NDDO)	SM20	5.70	6.15	7.00	6.99	5.74
Exact (NDDO)	SM21	4.10	3.34	6.16	4.09	4.24
Exact (NDDO)	SM22	2.40	0.30	2.70	1.02	1.16
Exact (NDDO)	SM22	7.43	4.61	5.93	5.52	5.36
Exact (NDDO)	SM23	5.45	4.25	7.45	4.96	6.76
Exact (NDDO)	SM24	2.60	2.03	5.10	2.69	4.12

Table 81.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	6.39	5.39	6.39	6.09
Point charge	SM02	5.03	4.16	3.45	4.16	4.99
Point charge	SM03	7.02	5.69	4.80	5.69	4.65
Point charge	SM04	6.02	5.32	4.59	5.30	6.12
Point charge	SM05	4.59	4.53	3.98	4.54	5.29
Point charge	SM06	3.03	3.21	2.53	3.22	3.34
Point charge	SM06	11.74	10.82	9.90	10.84	10.71
Point charge	SM07	6.08	5.43	4.68	5.38	6.11
Point charge	SM08	4.22	3.43	2.41	3.39	2.68
Point charge	SM09	5.37	3.80	3.06	3.77	4.52
Point charge	SM10	9.02	8.44	7.62	8.47	7.90
Point charge	SM11	3.89	3.15	2.48	3.12	3.70
Point charge	SM12	5.28	4.15	3.40	4.12	4.79
Point charge	SM13	5.77	4.47	3.72	4.41	5.47
Point charge	SM14	2.58	1.81	1.18	1.70	1.73
Point charge	SM14	5.30	4.33	3.70	4.28	4.87
Point charge	SM15	4.70	4.19	3.45	4.16	4.49
Point charge	SM15	8.94	6.65	5.69	6.67	6.34
Point charge	SM16	5.37	5.74	5.07	5.76	6.54
Point charge	SM16	10.65	10.63	9.72	10.63	9.84
Point charge	SM17	3.16	4.31	3.69	4.31	5.04
Point charge	SM18	2.15	2.45	1.81	2.44	3.57
Point charge	SM18	9.58	8.01	7.14	7.97	6.85
Point charge	SM18	11.02	8.74	7.74	8.76	8.76
Point charge	SM19	9.56	8.97	8.13	9.03	7.95
Point charge	SM20	5.70	6.19	5.34	6.20	4.76
Point charge	SM21	4.10	5.33	4.59	5.35	5.58
Point charge	SM22	2.40	3.56	2.75	3.60	3.83
Point charge	SM22	7.43	6.14	5.35	6.16	5.91
Point charge	SM23	5.45	4.18	3.45	4.16	6.18
Point charge	SM24	2.60	3.08	2.39	3.05	4.67

Table 82.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	5.58	6.46	6.19
Point charge	SM02	5.03	3.86	4.49	5.22
Point charge	SM03	7.02	5.06	5.84	4.93
Point charge	SM04	6.02	4.87	5.50	6.22
Point charge	SM05	4.59	4.34	4.82	5.49
Point charge	SM06	3.03	3.06	3.66	3.77
Point charge	SM06	11.74	9.55	10.38	10.26
Point charge	SM07	6.08	4.95	5.57	6.21
Point charge	SM08	4.22	2.95	3.81	3.19
Point charge	SM09	5.37	3.52	4.15	4.81
Point charge	SM10	9.02	7.54	8.29	7.79
Point charge	SM11	3.89	3.01	3.58	4.09
Point charge	SM12	5.28	3.83	4.46	5.05
Point charge	SM13	5.77	4.11	4.72	5.65
Point charge	SM14	2.58	1.87	2.33	2.35
Point charge	SM14	5.30	4.09	4.60	5.12
Point charge	SM15	4.70	3.87	4.50	4.79
Point charge	SM15	8.94	5.85	6.70	6.42
Point charge	SM16	5.37	5.30	5.90	6.59
Point charge	SM16	10.65	9.39	10.19	9.50
Point charge	SM17	3.16	4.08	4.62	5.27
Point charge	SM18	2.15	2.42	2.98	3.97
Point charge	SM18	9.58	7.12	7.85	6.87
Point charge	SM18	11.02	7.64	8.55	8.55
Point charge	SM19	9.56	7.99	8.78	7.84
Point charge	SM20	5.70	5.53	6.29	5.02
Point charge	SM21	4.10	4.87	5.54	5.74
Point charge	SM22	2.40	3.25	4.00	4.21
Point charge	SM22	7.43	5.54	6.26	6.04
Point charge	SM23	5.45	3.87	4.50	6.27
Point charge	SM24	2.60	2.93	3.51	4.94

Table 83.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/A level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on B3LYP-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.24.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	5.24	6.20	5.91
Point charge	SM02	5.03	3.36	4.04	4.85
Point charge	SM03	7.02	4.67	5.53	4.52
Point charge	SM04	6.02	4.46	5.15	5.94
Point charge	SM05	4.59	3.87	4.41	5.14
Point charge	SM06	3.03	2.47	3.13	3.25
Point charge	SM06	11.74	9.60	10.50	10.38
Point charge	SM07	6.08	4.55	5.23	5.93
Point charge	SM08	4.22	2.35	3.30	2.62
Point charge	SM09	5.37	2.98	3.67	4.39
Point charge	SM10	9.02	7.39	8.22	7.66
Point charge	SM11	3.89	2.42	3.04	3.61
Point charge	SM12	5.28	3.31	4.01	4.66
Point charge	SM13	5.77	3.63	4.29	5.32
Point charge	SM14	2.58	1.16	1.67	1.70
Point charge	SM14	5.30	3.60	4.17	4.73
Point charge	SM15	4.70	3.36	4.05	4.37
Point charge	SM15	8.94	5.53	6.47	6.16
Point charge	SM16	5.37	4.93	5.59	6.35
Point charge	SM16	10.65	9.42	10.30	9.54
Point charge	SM17	3.16	3.59	4.19	4.90
Point charge	SM18	2.15	1.77	2.38	3.47
Point charge	SM18	9.58	6.93	7.73	6.65
Point charge	SM18	11.02	7.50	8.50	8.49
Point charge	SM19	9.56	7.89	8.76	7.71
Point charge	SM20	5.70	5.19	6.02	4.62
Point charge	SM21	4.10	4.46	5.20	5.42
Point charge	SM22	2.40	2.68	3.51	3.73
Point charge	SM22	7.43	5.20	5.98	5.74
Point charge	SM23	5.45	3.36	4.05	6.00
Point charge	SM24	2.60	2.33	2.97	4.54

Table 84.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X reoptimised geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.93	11.12	9.91	9.43
Exact (Mult.)	SM02	5.03	3.76	5.24	3.73	5.09
Exact (Mult.)	SM03	7.02	8.67	9.96	8.64	6.99
Exact (Mult.)	SM04	6.02	4.93	6.38	4.86	6.18
Exact (Mult.)	SM05	4.59	2.63	4.20	2.47	3.68
Exact (Mult.)	SM06	3.03	1.17	2.68	1.18	1.35
Exact (Mult.)	SM06	11.74	11.11	12.39	11.14	10.95
Exact (Mult.)	SM07	6.08	4.80	6.27	4.78	5.94
Exact (Mult.)	SM08	4.22	5.20	6.36	5.14	4.03
Exact (Mult.)	SM09	5.37	3.85	5.32	3.80	5.02
Exact (Mult.)	SM10	9.02	9.46	10.75	9.40	8.49
Exact (Mult.)	SM11	3.89	2.98	4.53	2.95	3.91
Exact (Mult.)	SM12	5.28	3.93	5.41	3.90	5.00
Exact (Mult.)	SM13	5.77	3.61	5.06	3.54	5.26
Exact (Mult.)	SM14	2.58	0.57	2.21	0.44	0.49
Exact (Mult.)	SM14	5.30	4.21	5.77	4.14	5.10
Exact (Mult.)	SM15	4.70	3.53	4.98	3.48	4.01
Exact (Mult.)	SM15	8.94	9.50	10.73	9.51	8.99
Exact (Mult.)	SM16	5.37	4.26	5.81	4.27	5.51
Exact (Mult.)	SM16	10.65	11.57	12.85	11.59	10.35
Exact (Mult.)	SM17	3.16	2.36	3.97	2.37	3.53
Exact (Mult.)	SM18	2.15	1.96	3.49	1.83	3.67
Exact (Mult.)	SM18	9.58	9.10	10.51	9.08	7.29
Exact (Mult.)	SM18	11.02	10.77	11.84	10.77	10.77
Exact (Mult.)	SM19	9.56	10.50	11.83	10.56	8.86
Exact (Mult.)	SM20	5.70	6.67	8.03	6.71	4.34
Exact (Mult.)	SM21	4.10	3.67	5.16	3.72	4.06
Exact (Mult.)	SM22	2.40	0.66	2.31	0.69	1.03
Exact (Mult.)	SM22	7.43	7.13	8.34	7.22	6.84
Exact (Mult.)	SM23	5.45	4.63	6.07	4.56	7.84
Exact (Mult.)	SM24	2.60	1.45	2.96	1.40	4.02

Table 85.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.59	8.55	8.17
Exact (Mult.)	SM02	5.03	4.98	3.70	4.77
Exact (Mult.)	SM03	7.02	8.68	7.56	6.26
Exact (Mult.)	SM04	6.02	5.87	4.59	5.62
Exact (Mult.)	SM05	4.59	4.17	2.71	3.66
Exact (Mult.)	SM06	3.03	2.97	1.70	1.84
Exact (Mult.)	SM06	11.74	10.58	9.52	9.37
Exact (Mult.)	SM07	6.08	5.79	4.52	5.44
Exact (Mult.)	SM08	4.22	5.85	4.81	3.93
Exact (Mult.)	SM09	5.37	5.04	3.75	4.71
Exact (Mult.)	SM10	9.02	9.30	8.15	7.44
Exact (Mult.)	SM11	3.89	4.43	3.09	3.85
Exact (Mult.)	SM12	5.28	5.12	3.84	4.70
Exact (Mult.)	SM13	5.77	4.84	3.55	4.91
Exact (Mult.)	SM14	2.58	2.61	1.12	1.16
Exact (Mult.)	SM14	5.30	5.40	4.02	4.77
Exact (Mult.)	SM15	4.70	4.78	3.51	3.92
Exact (Mult.)	SM15	8.94	9.28	8.24	7.83
Exact (Mult.)	SM16	5.37	5.42	4.13	5.10
Exact (Mult.)	SM16	10.65	10.95	9.87	8.89
Exact (Mult.)	SM17	3.16	3.98	2.63	3.55
Exact (Mult.)	SM18	2.15	3.60	2.21	3.66
Exact (Mult.)	SM18	9.58	9.11	7.90	6.49
Exact (Mult.)	SM18	11.02	10.15	9.23	9.23
Exact (Mult.)	SM19	9.56	10.15	9.06	7.73
Exact (Mult.)	SM20	5.70	7.17	6.04	4.18
Exact (Mult.)	SM21	4.10	4.92	3.69	3.96
Exact (Mult.)	SM22	2.40	2.68	1.32	1.58
Exact (Mult.)	SM22	7.43	7.41	6.44	6.14
Exact (Mult.)	SM23	5.45	5.63	4.35	6.93
Exact (Mult.)	SM24	2.60	3.19	1.88	3.93

Table 86.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	9.89	8.66	8.22
Exact (Mult.)	SM02	5.03	4.45	2.94	4.20
Exact (Mult.)	SM03	7.02	8.82	7.49	5.96
Exact (Mult.)	SM04	6.02	5.50	3.98	5.20
Exact (Mult.)	SM05	4.59	3.49	1.77	2.89
Exact (Mult.)	SM06	3.03	2.07	0.57	0.74
Exact (Mult.)	SM06	11.74	11.06	9.81	9.63
Exact (Mult.)	SM07	6.08	5.40	3.91	4.99
Exact (Mult.)	SM08	4.22	5.48	4.24	3.21
Exact (Mult.)	SM09	5.37	4.52	3.00	4.13
Exact (Mult.)	SM10	9.02	9.55	8.19	7.35
Exact (Mult.)	SM11	3.89	3.79	2.22	3.11
Exact (Mult.)	SM12	5.28	4.61	3.10	4.11
Exact (Mult.)	SM13	5.77	4.28	2.76	4.36
Exact (Mult.)	SM14	2.58	1.64	-0.11	-0.07
Exact (Mult.)	SM14	5.30	4.94	3.32	4.20
Exact (Mult.)	SM15	4.70	4.21	2.71	3.20
Exact (Mult.)	SM15	8.94	9.53	8.29	7.81
Exact (Mult.)	SM16	5.37	4.97	3.44	4.58
Exact (Mult.)	SM16	10.65	11.49	10.22	9.07
Exact (Mult.)	SM17	3.16	3.27	1.67	2.75
Exact (Mult.)	SM18	2.15	2.82	1.18	2.89
Exact (Mult.)	SM18	9.58	9.33	7.89	6.23
Exact (Mult.)	SM18	11.02	10.56	9.46	9.46
Exact (Mult.)	SM19	9.56	10.55	9.26	7.69
Exact (Mult.)	SM20	5.70	7.03	5.70	3.50
Exact (Mult.)	SM21	4.10	4.37	2.93	3.24
Exact (Mult.)	SM22	2.40	1.73	0.12	0.44
Exact (Mult.)	SM22	7.43	7.31	6.17	5.82
Exact (Mult.)	SM23	5.45	5.22	3.70	6.75
Exact (Mult.)	SM24	2.60	2.33	0.78	3.20

Table 87.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	10.35	9.94	9.59
Exact (NDDO)	SM02	5.03	6.07	3.46	4.44
Exact (NDDO)	SM03	7.02	7.74	7.56	6.39
Exact (NDDO)	SM04	6.02	7.29	4.72	5.65
Exact (NDDO)	SM05	4.59	5.66	2.68	3.53
Exact (NDDO)	SM06	3.03	3.60	1.03	1.13
Exact (NDDO)	SM06	11.74	11.00	10.82	10.72
Exact (NDDO)	SM07	6.08	7.06	4.55	5.38
Exact (NDDO)	SM08	4.22	4.96	4.92	4.12
Exact (NDDO)	SM09	5.37	6.08	3.52	4.39
Exact (NDDO)	SM10	9.02	9.51	9.34	8.69
Exact (NDDO)	SM11	3.89	5.13	2.78	3.46
Exact (NDDO)	SM12	5.28	6.11	3.61	4.38
Exact (NDDO)	SM13	5.77	6.23	3.47	4.71
Exact (NDDO)	SM14	2.58	2.20	0.45	0.48
Exact (NDDO)	SM14	5.30	6.54	4.07	4.74
Exact (NDDO)	SM15	4.70	5.62	3.27	3.64
Exact (NDDO)	SM15	8.94	9.88	9.50	9.13
Exact (NDDO)	SM16	5.37	6.60	4.12	4.98
Exact (NDDO)	SM16	10.65	10.96	10.66	9.78
Exact (NDDO)	SM17	3.16	5.05	2.37	3.21
Exact (NDDO)	SM18	2.15	4.81	1.57	2.89
Exact (NDDO)	SM18	9.58	8.53	8.69	7.40
Exact (NDDO)	SM18	11.02	11.81	10.71	10.71
Exact (NDDO)	SM19	9.56	10.07	10.19	8.99
Exact (NDDO)	SM20	5.70	6.52	6.54	4.86
Exact (NDDO)	SM21	4.10	6.12	3.42	3.61
Exact (NDDO)	SM22	2.40	1.27	-0.91	-0.73
Exact (NDDO)	SM22	7.43	5.41	4.86	4.65
Exact (NDDO)	SM23	5.45	6.97	3.69	6.07
Exact (NDDO)	SM24	2.60	4.48	1.34	3.22

Table 88.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			/B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	8.60	9.85	9.53	9.26
Exact (NDDO)	SM02	5.03	3.88	6.58	4.58	5.33
Exact (NDDO)	SM03	7.02	6.86	7.85	7.71	6.82
Exact (NDDO)	SM04	6.02	4.84	7.51	5.54	6.25
Exact (NDDO)	SM05	4.59	3.44	6.26	3.98	4.63
Exact (NDDO)	SM06	3.03	2.01	4.69	2.73	2.80
Exact (NDDO)	SM06	11.74	9.30	10.34	10.21	10.13
Exact (NDDO)	SM07	6.08	4.70	7.33	5.42	6.05
Exact (NDDO)	SM08	4.22	4.78	5.73	5.70	5.09
Exact (NDDO)	SM09	5.37	3.93	6.58	4.63	5.29
Exact (NDDO)	SM10	9.02	8.25	9.20	9.08	8.58
Exact (NDDO)	SM11	3.89	3.41	5.86	4.06	4.58
Exact (NDDO)	SM12	5.28	4.00	6.60	4.70	5.29
Exact (NDDO)	SM13	5.77	3.90	6.70	4.59	5.54
Exact (NDDO)	SM14	2.58	1.78	3.62	2.28	2.30
Exact (NDDO)	SM14	5.30	4.43	6.94	5.05	5.56
Exact (NDDO)	SM15	4.70	3.72	6.23	4.43	4.72
Exact (NDDO)	SM15	8.94	8.28	9.49	9.20	8.92
Exact (NDDO)	SM16	5.37	4.40	6.98	5.08	5.74
Exact (NDDO)	SM16	10.65	9.19	10.31	10.08	9.41
Exact (NDDO)	SM17	3.16	3.12	5.80	3.75	4.39
Exact (NDDO)	SM18	2.15	2.54	5.61	3.14	4.15
Exact (NDDO)	SM18	9.58	7.82	8.45	8.58	7.59
Exact (NDDO)	SM18	11.02	9.08	10.96	10.12	10.12
Exact (NDDO)	SM19	9.56	8.85	9.63	9.72	8.81
Exact (NDDO)	SM20	5.70	6.09	6.92	6.94	5.65
Exact (NDDO)	SM21	4.10	3.79	6.61	4.55	4.69
Exact (NDDO)	SM22	2.40	0.52	2.91	1.24	1.38
Exact (NDDO)	SM22	7.43	4.75	6.07	5.65	5.49
Exact (NDDO)	SM23	5.45	4.07	7.27	4.76	6.57
Exact (NDDO)	SM24	2.60	2.29	5.36	2.96	4.40

Table 89.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			/B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	8.42	9.67	9.35	9.09
Exact (NDDO)	SM02	5.03	3.69	6.39	4.39	5.15
Exact (NDDO)	SM03	7.02	6.68	7.67	7.53	6.64
Exact (NDDO)	SM04	6.02	4.65	7.33	5.36	6.07
Exact (NDDO)	SM05	4.59	3.25	6.08	3.79	4.44
Exact (NDDO)	SM06	3.03	1.82	4.50	2.53	2.60
Exact (NDDO)	SM06	11.74	9.13	10.17	10.03	9.96
Exact (NDDO)	SM07	6.08	4.51	7.15	5.23	5.86
Exact (NDDO)	SM08	4.22	4.59	5.54	5.51	4.90
Exact (NDDO)	SM09	5.37	3.74	6.40	4.44	5.11
Exact (NDDO)	SM10	9.02	8.07	9.02	8.90	8.40
Exact (NDDO)	SM11	3.89	3.22	5.67	3.87	4.39
Exact (NDDO)	SM12	5.28	3.80	6.42	4.51	5.10
Exact (NDDO)	SM13	5.77	3.71	6.52	4.40	5.35
Exact (NDDO)	SM14	2.58	1.58	3.42	2.09	2.11
Exact (NDDO)	SM14	5.30	4.24	6.75	4.86	5.37
Exact (NDDO)	SM15	4.70	3.53	6.04	4.24	4.53
Exact (NDDO)	SM15	8.94	8.10	9.31	9.02	8.74
Exact (NDDO)	SM16	5.37	4.21	6.80	4.89	5.56
Exact (NDDO)	SM16	10.65	9.01	10.14	9.91	9.24
Exact (NDDO)	SM17	3.16	2.92	5.61	3.56	4.20
Exact (NDDO)	SM18	2.15	2.34	5.42	2.95	3.96
Exact (NDDO)	SM18	9.58	7.64	8.27	8.40	7.41
Exact (NDDO)	SM18	11.02	8.90	10.79	9.95	9.95
Exact (NDDO)	SM19	9.56	8.67	9.46	9.55	8.63
Exact (NDDO)	SM20	5.70	5.90	6.74	6.75	5.46
Exact (NDDO)	SM21	4.10	3.60	6.43	4.36	4.50
Exact (NDDO)	SM22	2.40	0.31	2.71	1.04	1.18
Exact (NDDO)	SM22	7.43	4.56	5.88	5.46	5.30
Exact (NDDO)	SM23	5.45	3.87	7.08	4.57	6.39
Exact (NDDO)	SM24	2.60	2.10	5.17	2.77	4.21

Table 90.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	6.43	5.27	6.42	6.12
Point charge	SM02	5.03	4.27	3.37	4.24	5.08
Point charge	SM03	7.02	5.78	4.71	5.75	4.71
Point charge	SM04	6.02	5.54	4.65	5.52	6.33
Point charge	SM05	4.59	3.63	2.80	3.48	4.24
Point charge	SM06	3.03	3.20	2.35	3.20	3.32
Point charge	SM06	11.74	10.15	9.09	10.16	10.04
Point charge	SM07	6.08	4.68	3.78	4.64	5.37
Point charge	SM08	4.22	3.48	2.30	3.42	2.72
Point charge	SM09	5.37	3.99	3.07	3.93	4.69
Point charge	SM10	9.02	7.85	6.81	7.80	7.23
Point charge	SM11	3.89	3.17	2.34	3.15	3.74
Point charge	SM12	5.28	4.31	3.40	4.27	4.95
Point charge	SM13	5.77	4.32	3.40	4.24	5.33
Point charge	SM14	2.58	1.70	0.88	1.55	1.59
Point charge	SM14	5.30	4.32	3.52	4.27	4.86
Point charge	SM15	4.70	4.20	3.30	4.16	4.49
Point charge	SM15	8.94	6.36	5.24	6.37	6.05
Point charge	SM16	5.37	5.80	4.97	5.81	6.59
Point charge	SM16	10.65	10.52	9.46	10.52	9.74
Point charge	SM17	3.16	4.24	3.46	4.25	4.97
Point charge	SM18	2.15	2.44	1.62	2.40	3.53
Point charge	SM18	9.58	7.95	6.96	7.94	6.82
Point charge	SM18	11.02	8.74	7.58	8.77	8.76
Point charge	SM19	9.56	8.48	7.48	8.53	7.47
Point charge	SM20	5.70	5.63	4.64	5.66	4.17
Point charge	SM21	4.10	5.55	4.63	5.56	5.79
Point charge	SM22	2.40	3.58	2.62	3.62	3.85
Point charge	SM22	7.43	6.10	5.13	6.12	5.88
Point charge	SM23	5.45	3.60	2.66	3.49	5.54
Point charge	SM24	2.60	3.17	2.32	3.14	4.78

Table 91.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.11 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	5.61	6.49	6.22
Point charge	SM02	5.03	3.94	4.56	5.31
Point charge	SM03	7.02	5.12	5.90	4.98
Point charge	SM04	6.02	5.07	5.69	6.41
Point charge	SM05	4.59	3.44	3.89	4.57
Point charge	SM06	3.03	3.04	3.65	3.76
Point charge	SM06	11.74	8.97	9.78	9.67
Point charge	SM07	6.08	4.30	4.92	5.56
Point charge	SM08	4.22	2.99	3.84	3.22
Point charge	SM09	5.37	3.68	4.29	4.96
Point charge	SM10	9.02	6.97	7.70	7.20
Point charge	SM11	3.89	3.04	3.60	4.12
Point charge	SM12	5.28	3.97	4.59	5.19
Point charge	SM13	5.77	3.97	4.57	5.52
Point charge	SM14	2.58	1.74	2.19	2.23
Point charge	SM14	5.30	4.07	4.59	5.11
Point charge	SM15	4.70	3.87	4.50	4.79
Point charge	SM15	8.94	5.58	6.45	6.16
Point charge	SM16	5.37	5.35	5.94	6.63
Point charge	SM16	10.65	9.30	10.10	9.41
Point charge	SM17	3.16	4.02	4.57	5.21
Point charge	SM18	2.15	2.40	2.94	3.94
Point charge	SM18	9.58	7.10	7.83	6.84
Point charge	SM18	11.02	7.64	8.55	8.55
Point charge	SM19	9.56	7.56	8.35	7.41
Point charge	SM20	5.70	5.05	5.82	4.50
Point charge	SM21	4.10	5.05	5.73	5.93
Point charge	SM22	2.40	3.28	4.02	4.22
Point charge	SM22	7.43	5.49	6.22	6.01
Point charge	SM23	5.45	3.31	3.90	5.71
Point charge	SM24	2.60	3.02	3.60	5.04

Table 92.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.13 on ONIOM-PCM/X geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 5.25.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	5.27	6.23	5.94
Point charge	SM02	5.03	3.44	4.12	4.94
Point charge	SM03	7.02	4.73	5.58	4.58
Point charge	SM04	6.02	4.68	5.36	6.15
Point charge	SM05	4.59	2.89	3.39	4.13
Point charge	SM06	3.03	2.45	3.12	3.24
Point charge	SM06	11.74	8.96	9.85	9.73
Point charge	SM07	6.08	3.83	4.51	5.21
Point charge	SM08	4.22	2.40	3.33	2.65
Point charge	SM09	5.37	3.15	3.83	4.56
Point charge	SM10	9.02	6.77	7.57	7.01
Point charge	SM11	3.89	2.45	3.07	3.64
Point charge	SM12	5.28	3.47	4.15	4.81
Point charge	SM13	5.77	3.47	4.13	5.18
Point charge	SM14	2.58	1.03	1.52	1.56
Point charge	SM14	5.30	3.59	4.15	4.72
Point charge	SM15	4.70	3.37	4.05	4.37
Point charge	SM15	8.94	5.24	6.19	5.88
Point charge	SM16	5.37	4.98	5.64	6.39
Point charge	SM16	10.65	9.32	10.20	9.44
Point charge	SM17	3.16	3.53	4.13	4.83
Point charge	SM18	2.15	1.74	2.34	3.44
Point charge	SM18	9.58	6.91	7.70	6.62
Point charge	SM18	11.02	7.50	8.50	8.49
Point charge	SM19	9.56	7.41	8.28	7.25
Point charge	SM20	5.70	4.66	5.50	4.06
Point charge	SM21	4.10	4.65	5.41	5.63
Point charge	SM22	2.40	2.71	3.52	3.74
Point charge	SM22	7.43	5.14	5.94	5.71
Point charge	SM23	5.45	2.75	3.40	5.38
Point charge	SM24	2.60	2.42	3.06	4.64

Table 93.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	10.75	11.97	10.74	10.26
Exact (Mult.)	SM02	5.03	2.82	4.28	2.80	4.15
Exact (Mult.)	SM03	7.02	13.99	15.13	14.04	12.39
Exact (Mult.)	SM04	6.02	3.83	5.27	3.77	5.08
Exact (Mult.)	SM05	4.59	3.46	5.11	3.44	4.66
Exact (Mult.)	SM06	3.03	0.98	2.47	0.99	1.17
Exact (Mult.)	SM06	11.74	10.17	11.45	10.19	10.01
Exact (Mult.)	SM07	6.08	3.78	5.24	3.77	4.93
Exact (Mult.)	SM08	4.22	3.44	4.63	3.45	2.32
Exact (Mult.)	SM09	5.37	2.95	4.40	2.91	4.12
Exact (Mult.)	SM10	9.02	11.75	13.21	11.80	10.87
Exact (Mult.)	SM11	3.89	1.78	3.31	1.76	2.71
Exact (Mult.)	SM12	5.28	3.00	4.46	2.96	4.05
Exact (Mult.)	SM13	5.77	2.73	4.18	2.69	4.40
Exact (Mult.)	SM14	2.58	0.16	1.71	0.04	0.08
Exact (Mult.)	SM14	5.30	3.94	5.62	3.92	4.85
Exact (Mult.)	SM15	4.70	2.93	4.36	2.87	3.40
Exact (Mult.)	SM15	8.94	10.20	11.45	10.20	9.70
Exact (Mult.)	SM16	5.37	3.38	4.89	3.40	4.63
Exact (Mult.)	SM16	10.65	10.84	12.11	10.86	9.62
Exact (Mult.)	SM17	3.16	1.46	3.06	1.49	2.65
Exact (Mult.)	SM18	2.15	3.60	4.90	3.62	5.40
Exact (Mult.)	SM18	9.58	8.05	9.49	8.02	6.24
Exact (Mult.)	SM18	11.02	12.51	13.52	12.51	12.51
Exact (Mult.)	SM19	9.56	13.55	14.70	13.56	11.87
Exact (Mult.)	SM20	5.70	7.58	8.96	7.65	5.31
Exact (Mult.)	SM21	4.10	2.64	4.09	2.69	3.03
Exact (Mult.)	SM22	2.40	0.10	1.71	0.12	0.46
Exact (Mult.)	SM22	7.43	6.78	8.04	6.87	6.49
Exact (Mult.)	SM23	5.45	3.52	4.96	3.48	6.73
Exact (Mult.)	SM24	2.60	1.14	2.59	1.06	3.68

Table 94.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	13.24	9.71	9.35
Exact (Mult.)	SM02	5.03	7.47	3.67	4.70
Exact (Mult.)	SM03	7.02	15.62	12.22	10.97
Exact (Mult.)	SM04	6.02	8.21	4.41	5.41
Exact (Mult.)	SM05	4.59	8.15	4.16	5.09
Exact (Mult.)	SM06	3.03	6.10	2.30	2.43
Exact (Mult.)	SM06	11.74	12.87	9.29	9.15
Exact (Mult.)	SM07	6.08	8.20	4.41	5.29
Exact (Mult.)	SM08	4.22	7.66	4.17	3.31
Exact (Mult.)	SM09	5.37	7.56	3.75	4.68
Exact (Mult.)	SM10	9.02	14.25	10.52	9.81
Exact (Mult.)	SM11	3.89	6.75	2.88	3.61
Exact (Mult.)	SM12	5.28	7.60	3.80	4.63
Exact (Mult.)	SM13	5.77	7.39	3.59	4.89
Exact (Mult.)	SM14	2.58	5.54	1.58	1.61
Exact (Mult.)	SM14	5.30	8.55	4.52	5.23
Exact (Mult.)	SM15	4.70	7.51	3.73	4.13
Exact (Mult.)	SM15	8.94	12.86	9.30	8.92
Exact (Mult.)	SM16	5.37	7.94	4.13	5.06
Exact (Mult.)	SM16	10.65	13.36	9.81	8.86
Exact (Mult.)	SM17	3.16	6.58	2.68	3.56
Exact (Mult.)	SM18	2.15	7.89	4.30	5.65
Exact (Mult.)	SM18	9.58	11.42	7.64	6.29
Exact (Mult.)	SM18	11.02	14.36	11.05	11.05
Exact (Mult.)	SM19	9.56	15.30	11.86	10.57
Exact (Mult.)	SM20	5.70	11.00	7.36	5.59
Exact (Mult.)	SM21	4.10	7.32	3.59	3.85
Exact (Mult.)	SM22	2.40	5.56	1.63	1.89
Exact (Mult.)	SM22	7.43	10.27	6.77	6.48
Exact (Mult.)	SM23	5.45	7.98	4.19	6.66
Exact (Mult.)	SM24	2.60	6.18	2.35	4.34

Table 95.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the multipole-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (Mult.)	SM01	9.53	10.92	9.87	9.45
Exact (Mult.)	SM02	5.03	4.33	3.05	4.21
Exact (Mult.)	SM03	7.02	13.63	12.69	11.28
Exact (Mult.)	SM04	6.02	5.17	3.89	5.01
Exact (Mult.)	SM05	4.59	5.03	3.60	4.65
Exact (Mult.)	SM06	3.03	2.76	1.50	1.65
Exact (Mult.)	SM06	11.74	10.48	9.39	9.24
Exact (Mult.)	SM07	6.08	5.15	3.88	4.88
Exact (Mult.)	SM08	4.22	4.62	3.61	2.64
Exact (Mult.)	SM09	5.37	4.43	3.14	4.18
Exact (Mult.)	SM10	9.02	11.98	10.78	9.97
Exact (Mult.)	SM11	3.89	3.49	2.16	2.98
Exact (Mult.)	SM12	5.28	4.47	3.19	4.13
Exact (Mult.)	SM13	5.77	4.24	2.96	4.43
Exact (Mult.)	SM14	2.58	2.12	0.69	0.72
Exact (Mult.)	SM14	5.30	5.47	4.01	4.81
Exact (Mult.)	SM15	4.70	4.39	3.11	3.57
Exact (Mult.)	SM15	8.94	10.47	9.41	8.97
Exact (Mult.)	SM16	5.37	4.84	3.56	4.62
Exact (Mult.)	SM16	10.65	11.04	9.97	8.91
Exact (Mult.)	SM17	3.16	3.27	1.93	2.92
Exact (Mult.)	SM18	2.15	4.86	3.75	5.28
Exact (Mult.)	SM18	9.58	8.79	7.53	6.01
Exact (Mult.)	SM18	11.02	12.25	11.38	11.38
Exact (Mult.)	SM19	9.56	13.26	12.29	10.84
Exact (Mult.)	SM20	5.70	8.34	7.21	5.21
Exact (Mult.)	SM21	4.10	4.16	2.96	3.25
Exact (Mult.)	SM22	2.40	2.12	0.75	1.04
Exact (Mult.)	SM22	7.43	7.55	6.54	6.22
Exact (Mult.)	SM23	5.45	4.91	3.63	6.42
Exact (Mult.)	SM24	2.60	2.87	1.56	3.80

Table 96.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	10.54	10.78	10.43
Exact (NDDO)	SM02	5.03	4.89	2.49	3.47
Exact (NDDO)	SM03	7.02	13.00	13.80	12.64
Exact (NDDO)	SM04	6.02	6.00	3.65	4.57
Exact (NDDO)	SM05	4.59	6.29	3.51	4.38
Exact (NDDO)	SM06	3.03	3.26	0.91	1.00
Exact (NDDO)	SM06	11.74	9.34	9.91	9.81
Exact (NDDO)	SM07	6.08	5.82	3.55	4.38
Exact (NDDO)	SM08	4.22	2.63	3.42	2.61
Exact (NDDO)	SM09	5.37	4.93	2.60	3.47
Exact (NDDO)	SM10	9.02	11.09	11.69	11.02
Exact (NDDO)	SM11	3.89	3.63	1.58	2.26
Exact (NDDO)	SM12	5.28	4.91	2.66	3.43
Exact (NDDO)	SM13	5.77	5.19	2.61	3.84
Exact (NDDO)	SM14	2.58	1.18	0.03	0.05
Exact (NDDO)	SM14	5.30	6.12	3.84	4.51
Exact (NDDO)	SM15	4.70	4.76	2.68	3.06
Exact (NDDO)	SM15	8.94	9.92	10.22	9.85
Exact (NDDO)	SM16	5.37	5.46	3.24	4.11
Exact (NDDO)	SM16	10.65	9.53	9.97	9.09
Exact (NDDO)	SM17	3.16	3.86	1.41	2.24
Exact (NDDO)	SM18	2.15	6.41	3.59	4.88
Exact (NDDO)	SM18	9.58	6.68	7.70	6.42
Exact (NDDO)	SM18	11.02	12.93	12.34	12.34
Exact (NDDO)	SM19	9.56	12.34	13.44	12.24
Exact (NDDO)	SM20	5.70	6.78	7.64	5.97
Exact (NDDO)	SM21	4.10	4.94	2.41	2.60
Exact (NDDO)	SM22	2.40	0.70	-1.28	-1.10
Exact (NDDO)	SM22	7.43	4.24	4.44	4.23
Exact (NDDO)	SM23	5.45	5.84	2.61	4.97
Exact (NDDO)	SM24	2.60	4.05	1.00	2.89

Table 97.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			/B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	9.72	10.25	10.43	10.17
Exact (NDDO)	SM02	5.03	3.68	6.01	4.21	4.95
Exact (NDDO)	SM03	7.02	11.88	12.10	12.70	11.83
Exact (NDDO)	SM04	6.02	4.55	6.85	5.08	5.77
Exact (NDDO)	SM05	4.59	4.58	7.06	4.98	5.63
Exact (NDDO)	SM06	3.03	2.49	4.79	3.03	3.09
Exact (NDDO)	SM06	11.74	9.09	9.35	9.78	9.71
Exact (NDDO)	SM07	6.08	4.47	6.71	5.00	5.63
Exact (NDDO)	SM08	4.22	4.16	4.32	4.91	4.30
Exact (NDDO)	SM09	5.37	3.77	6.04	4.29	4.94
Exact (NDDO)	SM10	9.02	10.53	10.67	11.12	10.62
Exact (NDDO)	SM11	3.89	3.05	5.07	3.52	4.03
Exact (NDDO)	SM12	5.28	3.81	6.03	4.34	4.91
Exact (NDDO)	SM13	5.77	3.78	6.24	4.30	5.23
Exact (NDDO)	SM14	2.58	1.96	3.23	2.36	2.38
Exact (NDDO)	SM14	5.30	4.86	6.93	5.23	5.73
Exact (NDDO)	SM15	4.70	3.82	5.91	4.36	4.63
Exact (NDDO)	SM15	8.94	9.31	9.79	10.01	9.74
Exact (NDDO)	SM16	5.37	4.25	6.44	4.77	5.42
Exact (NDDO)	SM16	10.65	9.12	9.49	9.82	9.17
Exact (NDDO)	SM17	3.16	2.93	5.24	3.40	4.02
Exact (NDDO)	SM18	2.15	4.36	7.16	5.03	6.00
Exact (NDDO)	SM18	9.58	7.59	7.36	8.13	7.16
Exact (NDDO)	SM18	11.02	10.73	12.05	11.60	11.60
Exact (NDDO)	SM19	9.56	11.64	11.60	12.43	11.53
Exact (NDDO)	SM20	5.70	7.41	7.43	8.08	6.82
Exact (NDDO)	SM21	4.10	3.56	6.05	4.15	4.29
Exact (NDDO)	SM22	2.40	0.92	2.87	1.38	1.52
Exact (NDDO)	SM22	7.43	4.92	5.53	5.67	5.52
Exact (NDDO)	SM23	5.45	3.78	6.72	4.30	6.07
Exact (NDDO)	SM24	2.60	2.60	5.38	3.09	4.51

Table 98.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the NDDO-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential potential	Name molname	p$K_{a,exp}$ pKa_exp	p$K_{a,corr}$ /B@PFL (All)	/B@PFL (Indiv.,+)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Exact (NDDO)	SM01	9.53	9.53	10.06	10.24	9.98
Exact (NDDO)	SM02	5.03	3.54	5.85	4.06	4.80
Exact (NDDO)	SM03	7.02	11.68	11.89	12.49	11.62
Exact (NDDO)	SM04	6.02	4.40	6.68	4.92	5.62
Exact (NDDO)	SM05	4.59	4.43	6.89	4.82	5.47
Exact (NDDO)	SM06	3.03	2.35	4.63	2.89	2.95
Exact (NDDO)	SM06	11.74	8.90	9.17	9.59	9.52
Exact (NDDO)	SM07	6.08	4.32	6.55	4.85	5.47
Exact (NDDO)	SM08	4.22	4.01	4.17	4.76	4.15
Exact (NDDO)	SM09	5.37	3.63	5.88	4.14	4.79
Exact (NDDO)	SM10	9.02	10.34	10.47	10.92	10.42
Exact (NDDO)	SM11	3.89	2.91	4.91	3.38	3.89
Exact (NDDO)	SM12	5.28	3.67	5.87	4.19	4.76
Exact (NDDO)	SM13	5.77	3.63	6.07	4.15	5.07
Exact (NDDO)	SM14	2.58	1.83	3.09	2.23	2.24
Exact (NDDO)	SM14	5.30	4.71	6.76	5.07	5.57
Exact (NDDO)	SM15	4.70	3.68	5.75	4.21	4.48
Exact (NDDO)	SM15	8.94	9.12	9.60	9.82	9.55
Exact (NDDO)	SM16	5.37	4.10	6.27	4.62	5.27
Exact (NDDO)	SM16	10.65	8.93	9.31	9.63	8.98
Exact (NDDO)	SM17	3.16	2.79	5.08	3.26	3.88
Exact (NDDO)	SM18	2.15	4.21	6.99	4.88	5.84
Exact (NDDO)	SM18	9.58	7.41	7.18	7.95	6.99
Exact (NDDO)	SM18	11.02	10.53	11.85	11.40	11.40
Exact (NDDO)	SM19	9.56	11.44	11.40	12.22	11.33
Exact (NDDO)	SM20	5.70	7.24	7.26	7.90	6.65
Exact (NDDO)	SM21	4.10	3.42	5.88	4.00	4.14
Exact (NDDO)	SM22	2.40	0.80	2.73	1.25	1.39
Exact (NDDO)	SM22	7.43	4.77	5.37	5.51	5.36
Exact (NDDO)	SM23	5.45	3.63	6.56	4.15	5.91
Exact (NDDO)	SM24	2.60	2.46	5.22	2.95	4.36

Table 99.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and *hr*-MP2 pK_a correction parameters on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$			
			hr-MP2 (All)	/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	7.25	6.11	7.24	6.96
Point charge	SM02	5.03	3.48	2.56	3.44	4.23
Point charge	SM03	7.02	11.06	9.97	11.13	10.16
Point charge	SM04	6.02	4.83	3.91	4.80	5.56
Point charge	SM05	4.59	4.63	3.86	4.58	5.30
Point charge	SM06	3.03	3.16	2.31	3.17	3.28
Point charge	SM06	11.74	8.94	7.83	8.94	8.83
Point charge	SM07	6.08	3.97	3.06	3.95	4.62
Point charge	SM08	4.22	1.63	0.52	1.64	0.97
Point charge	SM09	5.37	3.14	2.21	3.09	3.79
Point charge	SM10	9.02	10.00	9.10	10.04	9.49
Point charge	SM11	3.89	2.23	1.38	2.21	2.76
Point charge	SM12	5.28	3.50	2.58	3.47	4.10
Point charge	SM13	5.77	3.68	2.76	3.62	4.63
Point charge	SM14	2.58	1.21	0.35	1.08	1.11
Point charge	SM14	5.30	4.23	3.48	4.21	4.74
Point charge	SM15	4.70	3.77	2.85	3.73	4.04
Point charge	SM15	8.94	7.06	5.95	7.07	6.78
Point charge	SM16	5.37	5.00	4.13	5.01	5.73
Point charge	SM16	10.65	9.62	8.55	9.63	8.90
Point charge	SM17	3.16	3.39	2.59	3.41	4.08
Point charge	SM18	2.15	4.81	3.85	4.83	5.87
Point charge	SM18	9.58	7.76	6.78	7.74	6.71
Point charge	SM18	11.02	10.29	9.08	10.32	10.31
Point charge	SM19	9.56	12.66	11.56	12.67	11.69
Point charge	SM20	5.70	6.72	5.74	6.77	5.41
Point charge	SM21	4.10	4.51	3.54	4.52	4.73
Point charge	SM22	2.40	3.06	2.19	3.09	3.31
Point charge	SM22	7.43	5.75	4.69	5.76	5.54
Point charge	SM23	5.45	2.96	2.03	2.90	4.79
Point charge	SM24	2.60	2.80	1.86	2.73	4.26

Table 100.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters from table 5.10 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	6.84	7.82	7.56
Point charge	SM02	5.03	3.78	4.54	5.27
Point charge	SM03	7.02	10.18	11.18	10.28
Point charge	SM04	6.02	4.94	5.71	6.42
Point charge	SM05	4.59	4.89	5.52	6.19
Point charge	SM06	3.03	3.55	4.30	4.40
Point charge	SM06	11.74	8.33	9.29	9.18
Point charge	SM07	6.08	4.21	4.98	5.60
Point charge	SM08	4.22	2.01	2.98	2.36
Point charge	SM09	5.37	3.47	4.23	4.89
Point charge	SM10	9.02	9.42	10.24	9.73
Point charge	SM11	3.89	2.76	3.47	3.98
Point charge	SM12	5.28	3.79	4.56	5.15
Point charge	SM13	5.77	3.94	4.69	5.63
Point charge	SM14	2.58	1.87	2.50	2.53
Point charge	SM14	5.30	4.57	5.20	5.70
Point charge	SM15	4.70	4.03	4.79	5.07
Point charge	SM15	8.94	6.71	7.68	7.40
Point charge	SM16	5.37	5.13	5.89	6.56
Point charge	SM16	10.65	8.95	9.88	9.20
Point charge	SM17	3.16	3.80	4.50	5.13
Point charge	SM18	2.15	4.89	5.74	6.70
Point charge	SM18	9.58	7.42	8.25	7.29
Point charge	SM18	11.02	9.41	10.48	10.47
Point charge	SM19	9.56	11.55	12.51	11.59
Point charge	SM20	5.70	6.53	7.42	6.14
Point charge	SM21	4.10	4.62	5.47	5.67
Point charge	SM22	2.40	3.46	4.24	4.43
Point charge	SM22	7.43	5.61	6.54	6.33
Point charge	SM23	5.45	3.32	4.07	5.83
Point charge	SM24	2.60	3.17	3.92	5.35

Table 101.: Predicted and experimental pK_a values for the SAMPL6 data set from single point calculations at the ONIOM-EC-RISM/X level of theory using the point-charge-based ESP and ONIOM-EC-RISM/B@PFL pK_a correction parameters without thiols from table 5.12 on PM6-PCM geometries. The predicted pK_a values are shown separately for each set of PMV parameters used. These values correspond to the results shown in table 15.

Potential	Name	p$K_{a,exp}$	p$K_{a,corr}$		
			/B@PFL (All)	*hr*-MP2 (All) & *lr*-PM6 (All)	*hr*-MP2 (All) & *lr*-PM6 (indiv,+)
Point charge	SM01	9.53	6.62	7.66	7.40
Point charge	SM02	5.03	3.35	4.16	4.89
Point charge	SM03	7.02	10.18	11.25	10.35
Point charge	SM04	6.02	4.60	5.41	6.11
Point charge	SM05	4.59	4.54	5.21	5.87
Point charge	SM06	3.03	3.11	3.91	4.01
Point charge	SM06	11.74	8.21	9.23	9.12
Point charge	SM07	6.08	3.81	4.63	5.25
Point charge	SM08	4.22	1.47	2.50	1.88
Point charge	SM09	5.37	3.03	3.84	4.49
Point charge	SM10	9.02	9.37	10.24	9.74
Point charge	SM11	3.89	2.26	3.03	3.53
Point charge	SM12	5.28	3.37	4.18	4.77
Point charge	SM13	5.77	3.53	4.33	5.25
Point charge	SM14	2.58	1.32	1.98	2.02
Point charge	SM14	5.30	4.20	4.87	5.36
Point charge	SM15	4.70	3.62	4.43	4.71
Point charge	SM15	8.94	6.47	7.51	7.24
Point charge	SM16	5.37	4.79	5.60	6.27
Point charge	SM16	10.65	8.87	9.87	9.19
Point charge	SM17	3.16	3.38	4.13	4.75
Point charge	SM18	2.15	4.54	5.44	6.40
Point charge	SM18	9.58	7.24	8.13	7.17
Point charge	SM18	11.02	9.36	10.50	10.50
Point charge	SM19	9.56	11.65	12.67	11.76
Point charge	SM20	5.70	6.28	7.23	5.98
Point charge	SM21	4.10	4.25	5.15	5.35
Point charge	SM22	2.40	3.01	3.84	4.04
Point charge	SM22	7.43	5.31	6.30	6.09
Point charge	SM23	5.45	2.86	3.66	5.40
Point charge	SM24	2.60	2.70	3.50	4.91

4. Prediction of acidity constants and chemical shifts for a GEAEG pentapeptide

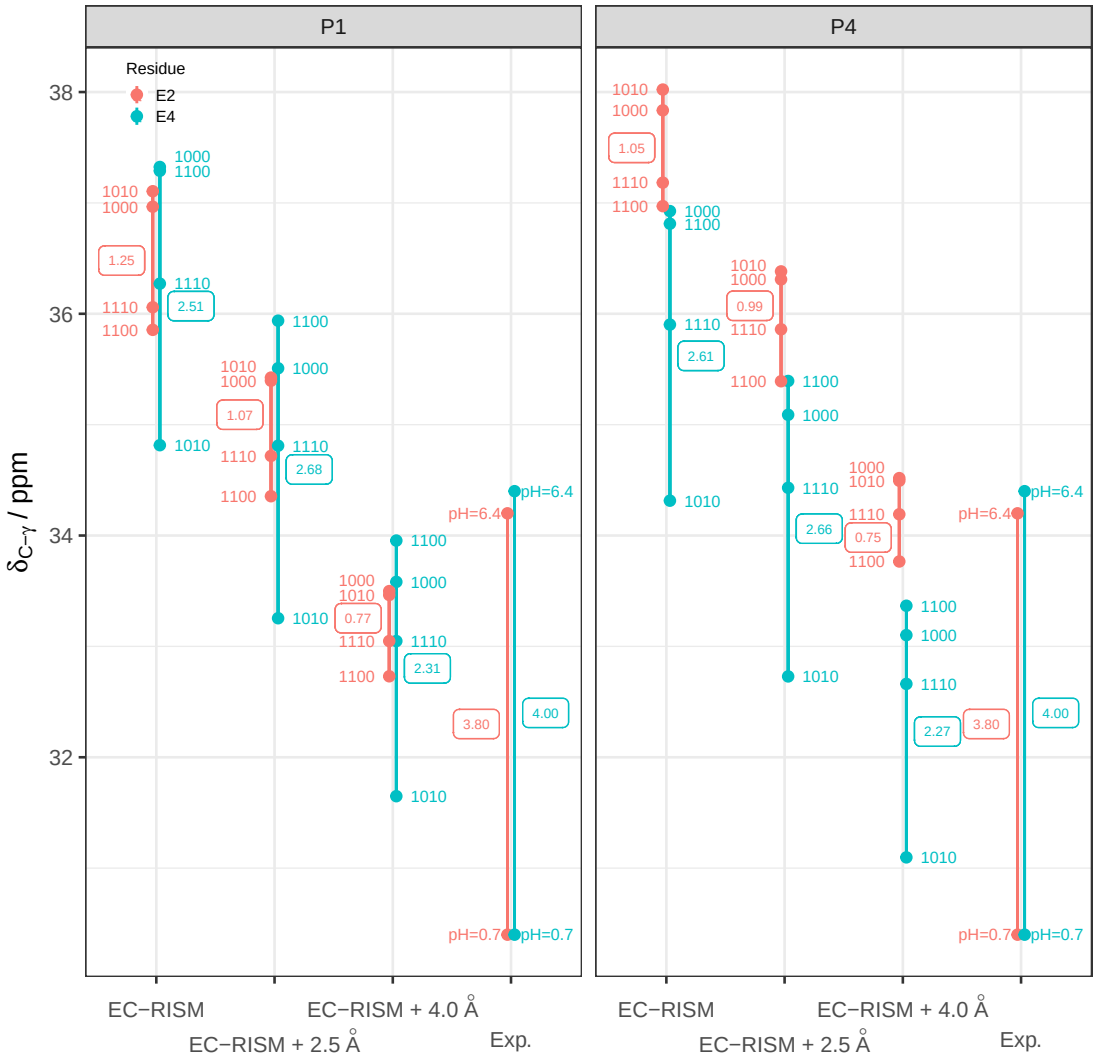

Figure 7.: Visualisation of the calculated chemical shifts with and without explicit water molecules in comparison with the experimental results. The points represent the averaged microstate chemical shifts or the end points of the NMR titration experiment, i.e. the measured shifts at pH 0.7 or 6.4. For the experiment, the lines represent the values that can be observed within the measured pH range. The lines connecting the calculated values represent the values that can be obtained for the pH-dependent weighted chemical shifts and can therefore be used to make an initial estimate of whether the method can reproduce the experiment. The values given in the boxes indicate the length of the lines and thus the difference between the maximum and minimum microstate shifts for the respective method or the range of shifts observed in the experiment.

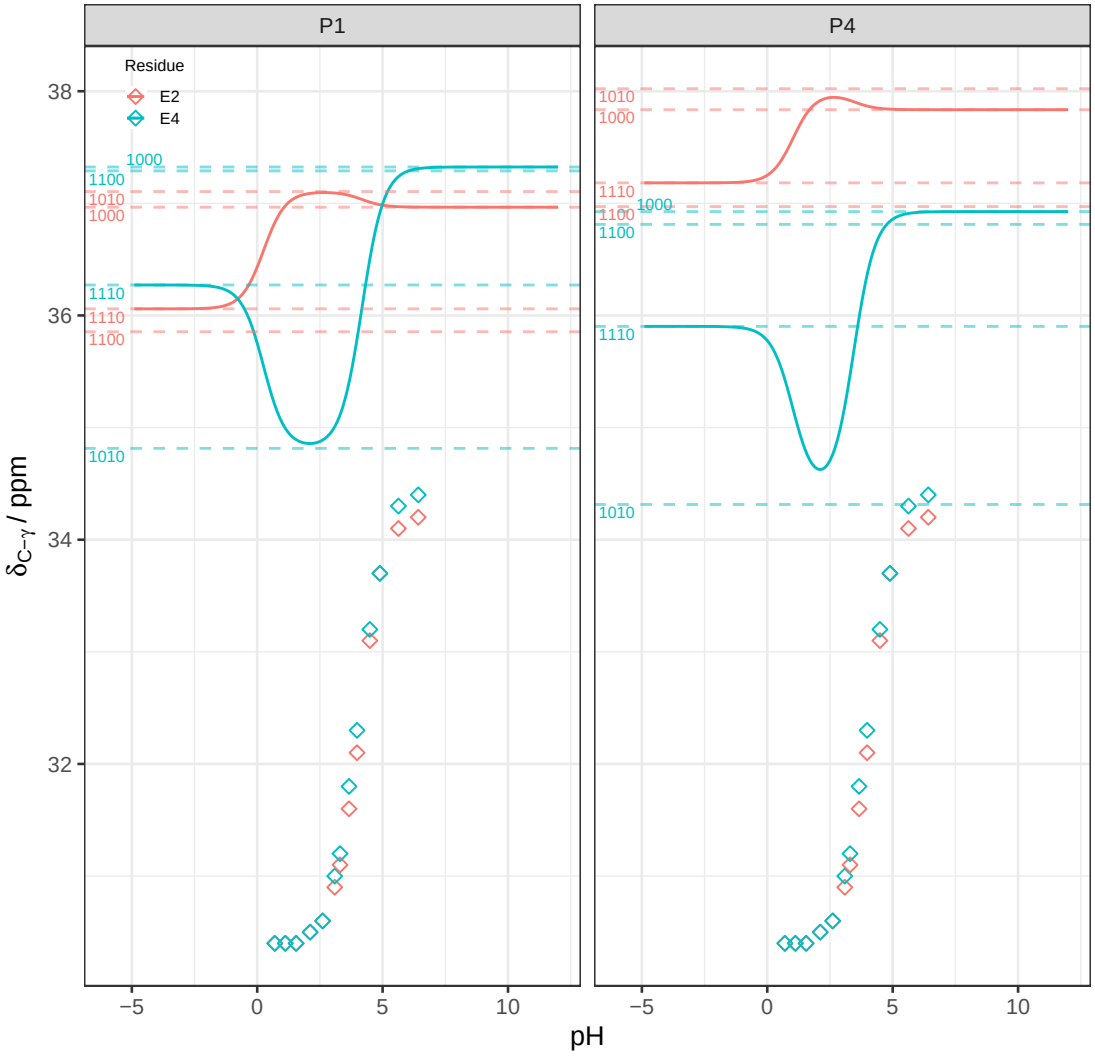

Figure 8.: Predicted pH-dependent chemical shift curves (equation 6.11) for the C_γ atoms of the GEAEG pentapeptide. The underlying populations were obtained using the /B@PFL pK_a correction. Experimental chemical shifts are shown as diamonds. The colours distinguish between the results obtained for either residue E2 (red) or E4 (blue). Dashed horizontal lines show the microstate shifts from table 6.6.

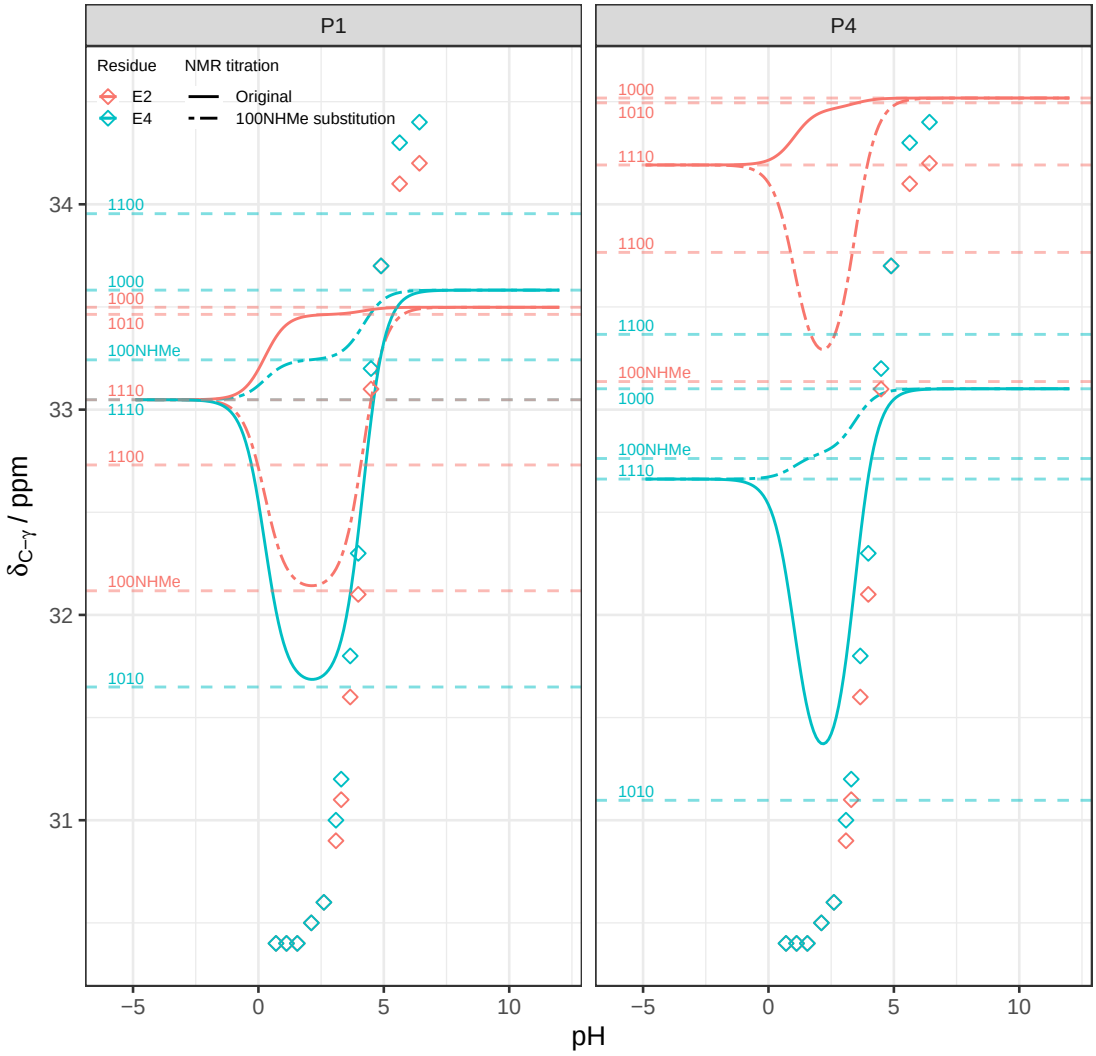

Figure 9.: Predicted pH-dependent chemical shift curves (equation 6.11) for the C_γ atoms of the GEAEG pentapeptide from ONIOM-EC-RISM/B calculations with an 4.0 Å explicit solvation shell. The underlying populations were obtained using the /B@PFL-pK_a correction. Experimental chemical shifts are shown as diamonds. The colours distinguish between the results obtained for either residue E2 (red) or E4 (blue). Dashed horizontal lines show the microstate shifts from tables 6.7 and 6.8

Table 102.: Change in pK_a for the given simulation time intervals in nanoseconds for all sampling rates and the *hr*-pK_a correction. The numerical values were calculated as the difference between the right and left values p$K_{a,corr,r}$ and p$K_{a,corr,l}$ of the given intervals. This is equal to the sum of all pK_a changes from one frame to the following and is used to estimate convergence of the pK_a for the complete set of MD simulation frames.

pK_a corr.	Part.	pK_a ID	f/ns^{-1}	p$K_{a,corr,r}$ − p$K_{a,corr,l}$						
				(50,100]	(100,150]	(150,200]	(200,250]	(250,300]	(300,350]	(350,400]
hr-MP2	P1	Macro pK_a 1	0.5	-1.57	-0.23	0.31	0.02	-0.20	-0.51	-0.07
hr-MP2	P1	Macro pK_a 1	1	-1.69	-0.29	0.14	-0.01	0.11	-0.38	0.01
hr-MP2	P1	Macro pK_a 1	2	-1.19	-0.06	0.13	-0.04	-0.02	-0.32	0.15
hr-MP2	P1	Macro pK_a 1	4	-1.38	-0.14	0.02	-0.02	-0.32	0.20	-0.13
hr-MP2	P1	Macro pK_a 2	0.5	-0.18	-0.93	-0.03	-0.21	-0.15	0.08	-0.02
hr-MP2	P1	Macro pK_a 2	1	0.28	-0.79	-0.15	-0.38	-0.45	-0.03	-0.19
hr-MP2	P1	Macro pK_a 2	2	0.03	-0.66	-0.06	-0.23	-0.38	-0.02	-0.27
hr-MP2	P1	Macro pK_a 2	4	0.02	-0.69	-0.09	-0.51	-0.07	-0.34	-0.03
hr-MP2	P1	1010 - 1000	0.5	0.20	-0.76	-0.03	-0.21	-0.15	0.08	-0.02
hr-MP2	P1	1010 - 1000	1	0.31	-0.78	-0.15	-0.38	-0.45	-0.03	-0.19
hr-MP2	P1	1010 - 1000	2	0.11	-0.63	-0.05	-0.23	-0.38	-0.02	-0.27
hr-MP2	P1	1010 - 1000	4	0.10	-0.66	-0.09	-0.51	-0.07	-0.34	-0.03
hr-MP2	P1	1100 - 1000	0.5	-0.39	-1.89	-0.61	-0.04	-0.28	-0.45	-0.29
hr-MP2	P1	1100 - 1000	1	-0.14	-1.46	-0.54	-0.29	-0.42	-0.32	-0.39
hr-MP2	P1	1100 - 1000	2	-0.31	-1.34	-0.59	-0.16	-0.43	-0.33	-0.31
hr-MP2	P1	1100 - 1000	4	-0.44	-1.38	-0.64	-0.55	-0.35	-0.28	0.01
hr-MP2	P1	1110 - 1010	0.5	-1.95	-0.40	0.30	0.02	-0.20	-0.51	-0.07
hr-MP2	P1	1110 - 1010	1	-1.72	-0.30	0.14	-0.01	0.11	-0.38	0.01
hr-MP2	P1	1110 - 1010	2	-1.28	-0.10	0.12	-0.04	-0.02	-0.32	0.15
hr-MP2	P1	1110 - 1010	4	-1.47	-0.16	0.02	-0.02	-0.32	0.20	-0.13
hr-MP2	P1	1110 - 1100	0.5	-1.36	0.73	0.89	-0.15	-0.07	0.01	0.20
hr-MP2	P1	1110 - 1100	1	-1.27	0.38	0.53	-0.11	0.08	-0.08	0.21
hr-MP2	P1	1110 - 1100	2	-0.86	0.62	0.66	-0.12	0.03	-0.01	0.19
hr-MP2	P1	1110 - 1100	4	-0.93	0.56	0.56	0.02	-0.04	0.14	-0.17
hr-MP2	P4	Macro pK_a 1	0.5	-0.96	0.51	0.54	0.10	-0.06	-0.35	-0.15
hr-MP2	P4	Macro pK_a 1	1	-1.45	0.01	0.18	-0.08	0.22	-0.31	-0.02
hr-MP2	P4	Macro pK_a 1	2	-1.40	0.37	0.37	-0.10	0.12	-0.23	0.16
hr-MP2	P4	Macro pK_a 1	4	-1.37	0.23	0.17	-0.24	0.15	-0.21	0.17
hr-MP2	P4	Macro pK_a 2	0.5	-0.39	-1.26	0.01	-0.19	-0.08	-0.01	0.13
hr-MP2	P4	Macro pK_a 2	1	-0.27	-0.89	-0.06	-0.22	-0.29	-0.02	-0.07
hr-MP2	P4	Macro pK_a 2	2	-0.06	-0.93	-0.11	-0.06	-0.21	-0.02	-0.15
hr-MP2	P4	Macro pK_a 2	4	0.03	-0.94	0.03	-0.01	-0.22	-0.02	-0.14
hr-MP2	P4	1010 - 1000	0.5	-0.16	-0.22	0.31	-0.19	-0.05	0.05	0.14
hr-MP2	P4	1010 - 1000	1	-0.30	-0.35	0.09	-0.19	-0.29	0.01	-0.07
hr-MP2	P4	1010 - 1000	2	-0.35	-0.28	0.18	-0.02	-0.20	0.01	-0.14
hr-MP2	P4	1010 - 1000	4	-0.30	-0.32	0.21	0.03	-0.22	-0.00	-0.14
hr-MP2	P4	1100 - 1000	0.5	-0.40	-1.36	-0.29	-0.18	-0.16	-0.49	-0.20
hr-MP2	P4	1100 - 1000	1	-0.27	-1.09	-0.40	-0.35	-0.26	-0.37	-0.33
hr-MP2	P4	1100 - 1000	2	-0.04	-1.03	-0.41	-0.20	-0.23	-0.32	-0.23
hr-MP2	P4	1100 - 1000	4	0.07	-1.11	-0.32	-0.25	-0.22	-0.26	-0.16
hr-MP2	P4	1110 - 1010	0.5	-1.20	-0.52	0.24	0.10	-0.09	-0.42	-0.16
hr-MP2	P4	1110 - 1010	1	-1.42	-0.53	0.03	-0.10	0.23	-0.34	-0.03
hr-MP2	P4	1110 - 1010	2	-1.11	-0.28	0.08	-0.14	0.12	-0.26	0.16
hr-MP2	P4	1110 - 1010	4	-1.05	-0.38	-0.01	-0.27	0.15	-0.22	0.17
hr-MP2	P4	1110 - 1100	0.5	-0.96	0.62	0.85	0.09	0.02	0.13	0.19
hr-MP2	P4	1110 - 1100	1	-1.45	0.21	0.52	0.06	0.19	0.03	0.24
hr-MP2	P4	1110 - 1100	2	-1.42	0.47	0.67	0.04	0.14	0.07	0.24
hr-MP2	P4	1110 - 1100	4	-1.42	0.41	0.52	0.00	0.15	0.03	0.20

Table 103.: Change in pK_a for the given simulation time intervals in nanoseconds for all sampling rates and the /B@PFL-pK_a correction. The numerical values were calculated as the difference between the right and left values p$K_{a,corr,r}$ and p$K_{a,corr,l}$ of the given intervals. This is equal to the sum of all pK_a changes from one frame to the following and is used to estimate convergence of the pK_a for the complete set of MD simulation frames.

pK_a corr.	Part.	pK_a ID	f/ns^{-1}	p$K_{a,corr,r}$ − p$K_{a,corr,l}$						
				(50,100]	(100,150]	(150,200]	(200,250]	(250,300]	(300,350]	(350,400]
/B@PFL	P1	Macro pK_a 1	0.5	-1.23	-0.18	0.24	0.02	-0.15	-0.40	-0.05
/B@PFL	P1	Macro pK_a 1	1	-1.33	-0.23	0.11	-0.01	0.09	-0.30	0.01
/B@PFL	P1	Macro pK_a 1	2	-0.93	-0.05	0.10	-0.03	-0.02	-0.25	0.12
/B@PFL	P1	Macro pK_a 1	4	-1.09	-0.11	0.01	-0.01	-0.25	0.16	-0.10
/B@PFL	P1	Macro pK_a 2	0.5	-0.14	-0.73	-0.03	-0.16	-0.12	0.06	-0.01
/B@PFL	P1	Macro pK_a 2	1	0.22	-0.62	-0.12	-0.30	-0.35	-0.02	-0.15
/B@PFL	P1	Macro pK_a 2	2	0.02	-0.52	-0.05	-0.18	-0.30	-0.02	-0.21
/B@PFL	P1	Macro pK_a 2	4	0.01	-0.54	-0.07	-0.40	-0.06	-0.27	-0.03
/B@PFL	P1	1010 - 1000	0.5	0.15	-0.59	-0.02	-0.16	-0.12	0.06	-0.01
/B@PFL	P1	1010 - 1000	1	0.24	-0.61	-0.12	-0.30	-0.35	-0.02	-0.15
/B@PFL	P1	1010 - 1000	2	0.09	-0.49	-0.04	-0.18	-0.30	-0.02	-0.21
/B@PFL	P1	1010 - 1000	4	0.08	-0.52	-0.07	-0.40	-0.06	-0.27	-0.03
/B@PFL	P1	1100 - 1000	0.5	-0.31	-1.48	-0.48	-0.03	-0.22	-0.35	-0.23
/B@PFL	P1	1100 - 1000	1	-0.11	-1.14	-0.42	-0.23	-0.33	-0.25	-0.31
/B@PFL	P1	1100 - 1000	2	-0.24	-1.05	-0.46	-0.13	-0.34	-0.26	-0.24
/B@PFL	P1	1100 - 1000	4	-0.34	-1.08	-0.50	-0.43	-0.28	-0.22	0.01
/B@PFL	P1	1110 - 1010	0.5	-1.53	-0.31	0.24	0.02	-0.15	-0.40	-0.05
/B@PFL	P1	1110 - 1010	1	-1.35	-0.24	0.11	-0.01	0.09	-0.30	0.01
/B@PFL	P1	1110 - 1010	2	-1.00	-0.08	0.10	-0.03	-0.02	-0.25	0.12
/B@PFL	P1	1110 - 1010	4	-1.15	-0.12	0.01	-0.01	-0.25	0.16	-0.10
/B@PFL	P1	1110 - 1100	0.5	-1.06	0.58	0.70	-0.12	-0.05	0.01	0.16
/B@PFL	P1	1110 - 1100	1	-1.00	0.30	0.41	-0.08	0.06	-0.07	0.16
/B@PFL	P1	1110 - 1100	2	-0.67	0.49	0.52	-0.09	0.02	-0.01	0.15
/B@PFL	P1	1110 - 1100	4	-0.73	0.44	0.44	0.02	-0.03	0.11	-0.14
/B@PFL	P4	Macro pK_a 1	0.5	-0.76	0.40	0.42	0.08	-0.05	-0.28	-0.12
/B@PFL	P4	Macro pK_a 1	1	-1.13	0.01	0.14	-0.06	0.17	-0.25	-0.02
/B@PFL	P4	Macro pK_a 1	2	-1.10	0.29	0.29	-0.08	0.09	-0.18	0.13
/B@PFL	P4	Macro pK_a 1	4	-1.08	0.18	0.14	-0.19	0.12	-0.16	0.13
/B@PFL	P4	Macro pK_a 2	0.5	-0.31	-0.98	0.01	-0.15	-0.06	-0.01	0.10
/B@PFL	P4	Macro pK_a 2	1	-0.21	-0.70	-0.05	-0.17	-0.22	-0.02	-0.06
/B@PFL	P4	Macro pK_a 2	2	-0.05	-0.73	-0.08	-0.04	-0.16	-0.02	-0.12
/B@PFL	P4	Macro pK_a 2	4	0.02	-0.73	0.02	-0.01	-0.17	-0.02	-0.11
/B@PFL	P4	1010 - 1000	0.5	-0.12	-0.17	0.24	-0.15	-0.04	0.04	0.11
/B@PFL	P4	1010 - 1000	1	-0.23	-0.28	0.07	-0.15	-0.23	0.00	-0.05
/B@PFL	P4	1010 - 1000	2	-0.28	-0.22	0.14	-0.01	-0.16	0.01	-0.11
/B@PFL	P4	1010 - 1000	4	-0.23	-0.25	0.16	0.02	-0.17	-0.00	-0.11
/B@PFL	P4	1100 - 1000	0.5	-0.31	-1.07	-0.23	-0.15	-0.13	-0.38	-0.16
/B@PFL	P4	1100 - 1000	1	-0.21	-0.86	-0.31	-0.27	-0.20	-0.29	-0.26
/B@PFL	P4	1100 - 1000	2	-0.03	-0.81	-0.32	-0.15	-0.18	-0.25	-0.18
/B@PFL	P4	1100 - 1000	4	0.05	-0.87	-0.25	-0.20	-0.17	-0.20	-0.13
/B@PFL	P4	1110 - 1010	0.5	-0.94	-0.41	0.19	0.08	-0.07	-0.33	-0.13
/B@PFL	P4	1110 - 1010	1	-1.11	-0.42	0.02	-0.08	0.18	-0.27	-0.02
/B@PFL	P4	1110 - 1010	2	-0.87	-0.22	0.06	-0.11	0.09	-0.21	0.12
/B@PFL	P4	1110 - 1010	4	-0.82	-0.30	-0.01	-0.21	0.12	-0.17	0.13
/B@PFL	P4	1110 - 1100	0.5	-0.76	0.48	0.66	0.07	0.02	0.10	0.15
/B@PFL	P4	1110 - 1100	1	-1.14	0.17	0.41	0.05	0.15	0.02	0.19
/B@PFL	P4	1110 - 1100	2	-1.12	0.37	0.53	0.03	0.11	0.05	0.19
/B@PFL	P4	1110 - 1100	4	-1.11	0.32	0.41	0.00	0.11	0.03	0.15

Overview of aids

The text of this thesis was typeset using LaTeX, more specifically pdfTeX in Overleaf (https://overleaf.com). Images of chemical structures were generated using ChemDraw20.1.1. Plots of data were generated using the R package "ggplot".[151] See the electronic supplementary material[116] for an overview of the software that was used to generate the data. Figures 2.1, 2.2, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 and 5.1 were generated using pdfTeX and PGF/TikZ.

The abstract given in the section "Kurzfassung / Abstract" was first written in German, then translated to English using the web version of DeepL (https://www.deepl.com/de/translator) and finally manually edited to give the final translated text.