Systematic QM Region Construction in QM/MM Calculations Based on Uncertainty Quantification

Felix Brandt and Christoph R. Jacob*

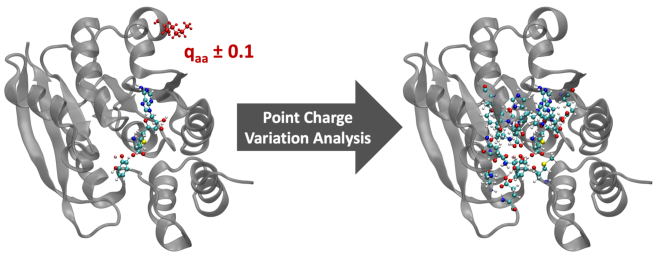
Institute of Physical and Theoretical Chemistry, Technische Universität Braunschweig, Gaußstr. 17, 38106 Braunschweig, Germany

E-mail: c.jacob@tu-braunschweig.de

Abstract

While QM/MM studies of enzymatic reactions are widely used in computational chemistry, the results of such studies are subject to numerous sources of uncertainty, and the effect of different choices by the simulation scientist that are required when setting up QM/MM calculations is often unclear. In particular, the selection of the QM region is crucial for obtaining accurate and reliable results. Simply including amino acids by their distance to the active site is mostly not sufficient as necessary residues are missing or unimportant residues are included without evidence. Here, we take a first step towards quantifying uncertainties in QM/MM calculations by assessing the sensitivity of QM/MM reaction energies with respect to variations of the MM point charges. We show that such a point charge variation analysis (PCVA) can be employed to judge the accuracy of QM/MM reaction energies obtained with a selected QM region, and devise a protocol to systematically construct QM regions that minimize this uncertainty. We apply such a PCVA to the example of catechol *O*-methyltransferase, and demonstrate that it provides a simple and reliable approach for the construction of the QM region. Our PCVA-based scheme is computationally efficient and requires only calculations for a system with a minimal QM region. Our work highlights the promise of applying methods of uncertainty quantification in computational chemistry.

Table of Contents Graphics



1 Introduction

Since the introduction of the QM/MM approach by Warshel and Levitt in 1976[1] it has evolved to a broadly used tool in computational chemistry. Dividing a large biomolecular system into two subsystems, one smaller subsystem treated quantum-mechanically (QM) and one larger subsystem treated with molecular mechanics (MM), allows one to investigate the mechanisms and energetics of enzymatic reactions in an efficient way.[2–4] Consequently, the application of the QM/MM approach to large biomolecules such as enzymes has become a common practice in the last two decades.[5–11]

However, in practice setting up QM/MM calculations is far from trivial and requires many manual choices by the simulation scientist.[4,8] Besides the selection of a suitable QM method and an MM force field, the most important decision is the choice of the QM region. The convergence of QM/MM results with the choice and particularly the size of the QM region has been investigated in several studies,[12–16] which underline that it generally has a large effect on the quality of the final results, and that often rather large QM regions are required for reaching converged results. To alleviate this problem, schemes for the systematic construction of QM regions have been proposed, usually with the aim of obtaining medium-sized QM regions that provide reliable QM/MM reaction energies. Examples of such schemes include free energy perturbation analysis,[17] charge deletion analysis,[18] charge shift analysis (CSA),[15] and Fukui shift analysis (FSA).[19] For their recently developed self-parametrizing system-focused atomistic models (SFAM), Brunken and Reiher proposed an automatic scheme for the construction of hybrid QM/SFAM models, including a systematic determination of the QM region based on the energy gradient.[20]

For a given QM region, there are different possible choices of the embedding method,[21] of a suitable coupling scheme,[22] and for the treatment of the boundary region[23] if covalent bonds cross the border between the subsystems. For the commonly used electrostatic embedding scheme,[4] the choice of the MM point charges will influence the resulting QM and QM/MM energies, and further parameters need to be chosen in advanced polarizable embedding[24,25] or

flexible embedding schemes.[26,27] Different coupling schemes are available such as IMOMM,[28] ONIOM,[29] or the AddRemove[30] model. The most common approach for treating covalent bonds across the QM–MM boundary is saturating the QM region with capping atoms.[4] Here, the position and the type of these atoms is crucial for the calculations. An alternative are frozen orbitals,[31] e.g., in the Localized SCF (LSCF)[32] or the Generalized Hybrid Orbital (GHO)[33] approach. The uncertainties introduced by all these different choices and the corresponding parameters are interconnected and will again depend on the choice of the QM region.

Consequently, there is a need for rigorous uncertainty quantification for QM/MM calculations, i.e., to systematically assess the sensitivity of the QM/MM energy with respect to these technical choices and empirical parameters and to ultimately provide rigorous error bounds on the results of QM/MM calculations (compared to a full QM treatment). Mathematical and computational tools for quantifying uncertainties in computer simulations have been developed intensively in the past decades (for textbooks, see, e.g. Refs. 34,35) and are employed in many areas of simulation science,[36,37] but their application is just starting to emerge in computational chemistry.[38] Recently, we have applied such tools for analyzing the sensitivity of calculated spectra with respect to distortions of the molecular structure.[39,40]

Here, we aim at taking a first step towards uncertainty quantification for QM/MM methods by analyzing the sensitivity of QM/MM reaction energies with respect to variations of the MM point charges. While there are other relevant empirical parameters entering in QM/MM calculations, most importantly those related to the treatment of the QM–MM boundary (e.g., to the placement of the link atoms), we expect the MM charges to be a key factor influencing the final QM/MM results.

The ability to quantify the sensitivity of QM/MM calculations with respect to parameters of the MM environment provides a natural starting point for guiding the systematic choice of the QM region. With increasing size of the QM region and approaching a full QM calculation, one can expect this sensitivity to decrease. Therefore, choosing the QM region such that the

uncertainty is reduced implies that the QM/MM calculation approach those of a full QM calculation. Here, we exploit this idea by proposing a simple and efficient scheme for the systematic construction of the QM region that is guided by uncertainty quantification.

This work is organized as follows. In Section 2 we recall the necessary theoretical background of QM/MM approaches (Sect. 2.1, introduce our point charge variation analysis (PCVA) for analyzing the sensitivity of QM/MM energies (Sect. 2.2, and give the computational details (Section 2.3). In Section 3, we introduce the model system used in this work and discuss the convergence of ligand charges and reaction energies for QM regions of increasing size. The sensitivity of these quantities with respect to global point charge variations is analyzed in Section 4. This is followed in by the evaluation of the energy sensitivity for single amino acids in Section 5, which are used to devise a scheme for the systematic construction of QM regions based on a PCVA. The QM regions obtained with this PCVA-based scheme are assessed for QM regions of increasing size and for atom-economical QM regions in Sections 6 and 7, respectively. Finally, conclusions and an outlook can be found in Section 8.

2 Methodology

2.1 QM/MM energy partitioning

QM/MM is based on the partitioning of the full target system into a QM region (A) including the interesting part of the system, such as the active center of an enzyme, and an MM region (B) containing all other atoms, i.e., the active center's environment. The total energy can be expressed as the sum of the QM energy of subsystem A, $E_{\mathrm{QM}}(\mathrm{A})$, the MM energy of subsystem B, $E_{\mathrm{MM}}(\mathrm{B})$, and an interaction energy between the two subsystems, $E_{\mathrm{int}}(\mathrm{A},\mathrm{B})$,

$$\begin{aligned} E_{\mathrm{QM/MM}} &= E_{\mathrm{QM}}(\mathrm{A}) + E_{\mathrm{MM}}(\mathrm{B}) + E_{\mathrm{int}}(\mathrm{A},\mathrm{B}) \\ &= E_{\mathrm{QM}}(\mathrm{A}) + E_{\mathrm{MM}}(\mathrm{B}) + E_{\mathrm{int,el}}(\mathrm{A},\mathrm{B}) + E_{\mathrm{int,ne}}(\mathrm{A},\mathrm{B}). \end{aligned} \tag{1}$$

The exact definition of these three energy contributions varies between different implementations of QM/MM schemes.[4,8] In the following, we focus on the general principles as far as they are relevant for the current work. Details on the QM/MM implementation employed here are given in Section 2.3.

The energy of the QM region, $E_{\mathrm{QM}}(\mathrm{A})$, is obtained from a quantum-chemical calculation of the corresponding subsystem A. To allow for covalent bonds to cross the boundary between the QM and MM regions, the QM subsystem is usually saturated using capping atoms. The energy of the MM region, $E_{\mathrm{MM}}(\mathrm{B})$, is calculated using a classical force field and contains the usual bonding and non-bonding force-field energy contributions. Finally, the interaction energy $E_{\mathrm{int}}(\mathrm{A},\mathrm{B})$ contains both electrostatic interactions $[E_{\mathrm{int,el}}(\mathrm{A})]$ and non-electrostatic interactions $[E_{\mathrm{int,ne}}(\mathrm{A},\mathrm{B})]$ between the two subsystems.

In the electrostatic embedding scheme, which is commonly used when applying QM/MM to enzymatic reactions, the electron density of subsystem A is polarized by the MM point charges of subsystem B. The MM point charges are included in the QM Hamiltonian[4] to compute the interaction between the electron density $\rho_{\mathrm{A}}(\boldsymbol{r})$ of subsystem A and the electrostatic potential $V_{\mathrm{B}}(\boldsymbol{r})$ derived from the point charges of subsystem B,

$$E_{\mathrm{int,el}}(\mathrm{A},\mathrm{B}) = \sum_{I=1}^{N_{\mathrm{B}}} \int \rho_{\mathrm{A}}(\boldsymbol{r}) \frac{q_{I,\mathrm{B}}}{|\boldsymbol{r} - \boldsymbol{R}_{I,\mathrm{B}}|}\, d^3r, \tag{2}$$

where N_B is the number of atoms in subsystem B and $q_{I,B}$ represents the MM point charge and $R_{i,B}$ the position of the I-th atom. This electrostatic interaction energy is usually included in the QM energy of subsystem A, i.e.,

$$E^{\mathrm{emb}}_{\mathrm{QM}}(\mathrm{A}, V_{\mathrm{B}}) = E_{\mathrm{QM}}(\mathrm{A}) + E_{\mathrm{int,el}}(\mathrm{A},\mathrm{B}). \tag{3}$$

Altogether, the QM/MM energy can be expressed as

$$E_{\mathrm{QM/MM}} = E^{\mathrm{emb}}_{\mathrm{QM}}(\mathrm{A}, V_{\mathrm{B}}) + E_{\mathrm{MM}}(\mathrm{B}) + E_{\mathrm{int,ne}}(\mathrm{A},\mathrm{B}). \tag{4}$$

It thus consists of the embedded QM energy of subsystem A, the MM energy of subsystem B, and the non-electrostatic interactions between the two subsystems.

When investigating enzymatic reactions with QM/MM calculations, the main quantity of interest (QoI) is generally the reaction energy,

$$\Delta E^{\text{reaction}}_{\text{QM/MM}} = E_{\text{QM/MM}}(\text{product}) - E_{\text{QM/MM}}(\text{reactant}). \tag{5}$$

2.2 Sensitivity analysis for QM/MM energies

The reaction energy calculated within a QM/MM model is subject to numerous sources of uncertainty (see Introduction). One important element of uncertainty quantification[34,35] is the analysis of the sensitivity of the simulation results with respect to its input parameters.[41] Here, we consider the QM/MM reaction energy $\Delta E^{\text{reaction}}_{\text{QM/MM}}$ as our quantity of interest (QoI) and analyze how sensitively it depends on parameters of the QM/MM model.

We consider the MM point charges $\boldsymbol{q}_{\text{MM}}$ as one of the most important sources of uncertainty and want to systematically analyze the effect of variations in the MM point charges on our QoI, i.e., the reaction energy $\Delta E^{\text{reaction}}_{\text{QM/MM}}(\boldsymbol{q}_{\text{MM}})$. To this end, we follow our earlier work on the sensitivity of calculated spectra with respect to distortions of the molecular structure[39] and consider a collective variation of the MM point charges, i.e.,

$$\boldsymbol{q}_{\text{MM}} = \boldsymbol{q}^0_{\text{MM}} + \Delta\boldsymbol{q}_{\text{MM}}(\Delta q), \tag{6}$$

where $\boldsymbol{q}_{\text{MM}}$ is a vector of size N_{B} containing all MM point charges, $\boldsymbol{q}^0_{\text{MM}}$ is the vector of the undistorted MM point charges as provided by the employed force field, and $\Delta\boldsymbol{q}_{\text{MM}}$ is a collective variation of these point charges, which depends on a parameter Δq that controls the size of the variation. We chose the collective variations of the MM point charges such that $\sum_I \Delta q_{\text{MM},I} = 0$, i.e., the sum of the MM point charges is preserved.

In the following, we will consider two types of collective point-charge variations. First,

we change the charges of all protein MM atoms simultaneously by an equal magnitude Δq, while changing all solvent MM charges equally such that the total charge is preserved, i.e,

$$\Delta q^{\mathrm{tot}}_{\mathrm{MM},I} = \begin{cases} +\Delta q & \text{for} \quad I \in \text{protein} \\ -\Delta q \cdot \left(N_{\mathrm{B}}^{\mathrm{protein}} / N_{\mathrm{B}}^{\mathrm{solvent}}\right) & \text{for} \quad I \in \text{solvent}, \end{cases} \tag{7}$$

where $N_{\mathrm{B}}^{\mathrm{protein}}$ and $N_{\mathrm{B}}^{\mathrm{solvent}}$ are the numbers of protein and solvent atoms in subsystem B, respectively. Second, we consider variations of the MM charges of the i-th amino acid,

$$\Delta q^{\mathrm{aa},i}_{\mathrm{MM},I} = \begin{cases} +\Delta q / N_{\mathrm{aa},i} & \text{for} \quad I \in \text{amino acid } i \\ -\Delta q / (N_{\mathrm{B}} - N_{\mathrm{aa},i}) & \text{for} \quad I \notin \text{amino acid } i, \end{cases} \tag{8}$$

where $N_{\mathrm{aa},i}$ is the number of atoms in the i-th amino acid. The first will provide an estimate of the overall sensitivity of the QM/MM reaction energy to variations of the MM point charges, while the second will allow us to assess the effect of the individual single amino acids.

For these collective point-charge variations, we perform a local sensitivity analysis [41] and consider the derivative of ΔE with respect to a the parameter Δq,

$$\delta \Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}} = \left.\frac{\partial \Delta E(\boldsymbol{q}_{\mathrm{MM}})}{\partial \Delta q}\right|_{\boldsymbol{q}^0_{\mathrm{MM}}} = \left.\frac{\partial \Delta E(\boldsymbol{q}^0_{\mathrm{MM}} + \Delta \boldsymbol{q}_{\mathrm{MM}}(\Delta q))}{\partial \Delta q}\right|_{\boldsymbol{q}^0_{\mathrm{MM}}}, \tag{9}$$

that is, the derivative is taken in the direction of the collective point-charge variation.

Of the components of the QM/MM energy [see Eq. (4)], the non-electrostatic interaction energy does not depend on the MM point charges. Therefore, the sensitivity of the QM/MM energy of the reactants or the products is given by,

$$\delta E_{\mathrm{QM/MM}} = \frac{\partial E^{\mathrm{emb}}_{\mathrm{QM}}(\mathrm{A}, V_{\mathrm{B}}(\boldsymbol{q}_{\mathrm{MM}}))}{\partial \Delta q} + \frac{\partial E_{\mathrm{MM}}(\mathrm{B}(\boldsymbol{q}_{\mathrm{MM}}))}{\partial \Delta q} \tag{10}$$

and the sensitivity of the reaction energy can be calculated as

$$\delta\Delta E^{\text{reaction}}_{\text{QM/MM}} = \left(\frac{\partial E^{\text{emb}}_{\text{QM}}(\text{A}^{\text{R}}, V^{\text{R}}_{\text{B}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q} - \frac{\partial E^{\text{emb}}_{\text{QM}}(\text{A}^{\text{P}}, V^{\text{P}}_{\text{B}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q} \right) + \left(\frac{\partial E_{\text{MM}}(\text{B}^{\text{R}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q} - \frac{\partial E_{\text{MM}}(\text{B}^{\text{P}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q} \right) \tag{11}$$

where the superscripts $^{\text{R}}$ and $^{\text{P}}$ designate the reactants and products, respectively. If the protein environment is similar for the reactants and the products, which is usually the case for enzymatic reactions, the second term can be expected to be small and could possibly neglected, i.e,

$$\delta\Delta E^{\text{reaction}}_{\text{QM/MM}} \approx \frac{\partial E^{\text{emb}}_{\text{QM}}(\text{A}^{\text{R}}, V^{\text{R}}_{\text{B}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q} - \frac{\partial E^{\text{emb}}_{\text{QM}}(\text{A}^{\text{P}}, V^{\text{P}}_{\text{B}}(\boldsymbol{q}_{\text{MM}}))}{\partial \Delta q}. \tag{12}$$

The simplest way of obtaining the derivatives necessary for the calculation of the sensitivity $\delta\Delta E^{\text{reaction}}_{\text{QM/MM}}$ is their numerical evaluation, either using a symmetric two-point finite difference formula,

$$\delta\Delta E^{\text{reaction}}_{\text{QM/MM}} = \left.\frac{\partial \Delta E(\boldsymbol{q}_{\text{MM}})}{\partial \Delta q}\right|_{\boldsymbol{q}^0_{\text{MM}}} \approx \frac{\Delta E(\boldsymbol{q}^0_{\text{MM}} + \Delta\boldsymbol{q}_{\text{MM}}) - \Delta E(\boldsymbol{q}^0_{\text{MM}} - \Delta\boldsymbol{q}_{\text{MM}})}{2\Delta q} \tag{13}$$

or using a forward two-point finite difference formula,

$$\delta\Delta E^{\text{reaction}}_{\text{QM/MM}} = \left.\frac{\partial \Delta E(\boldsymbol{q}_{\text{MM}})}{\partial \Delta \boldsymbol{q}}\right|_{\boldsymbol{q}^0_{\text{MM}}} \approx \frac{\Delta E(\boldsymbol{q}^0_{\text{MM}} + \Delta\boldsymbol{q}_{\text{MM}}) - \Delta E(\boldsymbol{q}^0_{\text{MM}})}{\Delta q}. \tag{14}$$

An analytical evaluation of these derivatives is also possible, but would require modifications of the quantum-chemical software packages used for subsystem A.

2.3 Computational Details

Molecular dynamics calculations were performed using GROMACS 2019.3[42,43] with the AMBER99SB-ILDN[44] force field. The already equilibrated initial structure provided by Ku-

lik *et al.* in the Supporting Information of Ref. 15 was solvated in TIP3P[45] water molecules in a cubic simulation box with 1 nm distance between the borders and the enzyme. The system was neutralized by adding six sodium cations. The positions of the solvent molecules and ions were minimized using the force field, while the enzyme structure was held fixed. Subsequently, a droplet was extracted including COMT with substrates, sodium ions and all water molecules within 33 Å from the COMT center of mass for the following QM/MM calculations.

All QM/MM calculations were performed using the Amsterdam Modeling Suite (AMS Version 2020.203).[46] The Amsterdam Density Functional (ADF) engine[47] was used for the QM part applying density functional theory (DFT) with the PBE exchange-correlation functional[48] employing a DZ and a TZP Slater-type orbital basis set[49] for all geometry optimizations and single point calculations, respectively. For the MM region the ForceField engine of AMS was used with the AMBER95 force field,[50] which was extended by parameters for SAM and catecholate using ANTECHAMBER[51,52] and ACPYPE.[53,54]

Electrostatic embedding as implemented in AMS[55] was applied for the interaction between the QM and MM regions. Link atoms were placed on the C_α–C and C_α–N bonds only including the α-carbon atom in the QM region for single QM amino acids, while also including the remaining backbone atoms between two subsequent QM amino acids to reduce the number of link atoms. Starting with QM region **2** and larger the water molecule which is located in the active site and which is resolved in the crystal structure, is included in the QM region. No other water molecules are considered for the QM part. Residues included in the different sized QM regions are listed in the Supporting Information in Tab. S1 with the corresponding QM region charge and the number of atoms and link atoms.

All QM/MM geometry optimizations were performed using the FIRE minimization algorithm[56] with all solvent molecules fixed to their initial coordinates. All charges evaluated for charge convergence tests are calculated from the Voronoi deformation density (VDD)[57] of the reactant structure only.

Modification of the input files concerning point charge variation and analysis of the results were achieved using Python. Plots were generated with MATPLOTLIB[58,59] and structures were visualized using VMD.[60]

A data set containing PDB files of the reactant and product starting structures, a modified AMBER95 force field file, AMS fragment files for the ligands and ions, and the AMS input files for all geometry optimizations and single point calculations is available at Ref. 61.

3 Catechol *O*-methyltransferase as model system for QM/MM calculations

As model system for investigating the sensitivity of QM/MM calculations on point-charge variations and for exploring automatic QM region selection schemes, we chose the enzyme catechol *O*-methyltransferase (COMT),[62] which plays a crucial role in the regulation of neurotransmitters in the human body. In a previous study Kulik *et al.*[15] investigated the convergence of catalytic properties with increasing QM region size and established it as a test case for benchmarking QM/MM approaches, in particular of schemes for the systematic determination of the QM region.[14,19,64]

We use the structural model of COMT introduced by Kulik *et al.*, starting from the initial equilibrated MM structure taken from the Supporting Information of Ref. 15 based on the protein crystal structure (PDB: 3BWM).[63] The COMT active site includes the neutral *S*-adenosyl methionine (SAM), the catecholate anion (CAT), and the catalytically active Mg^{2+} (see Fig. 1). After solvation and neutralization of the initial structure and a subsequent MM energy minimization of water molecules and ions, we performed QM/MM geometry optimizations with different sized QM regions (see Section 2.3). The different QM regions for our first convergence tests have been chosen based on the distance of the individual amino acid residues to the active site (see Table S1 in the Supporting Information). These QM regions match those used by Kulik *et al.* in Ref. 15. In our computational setup, only small

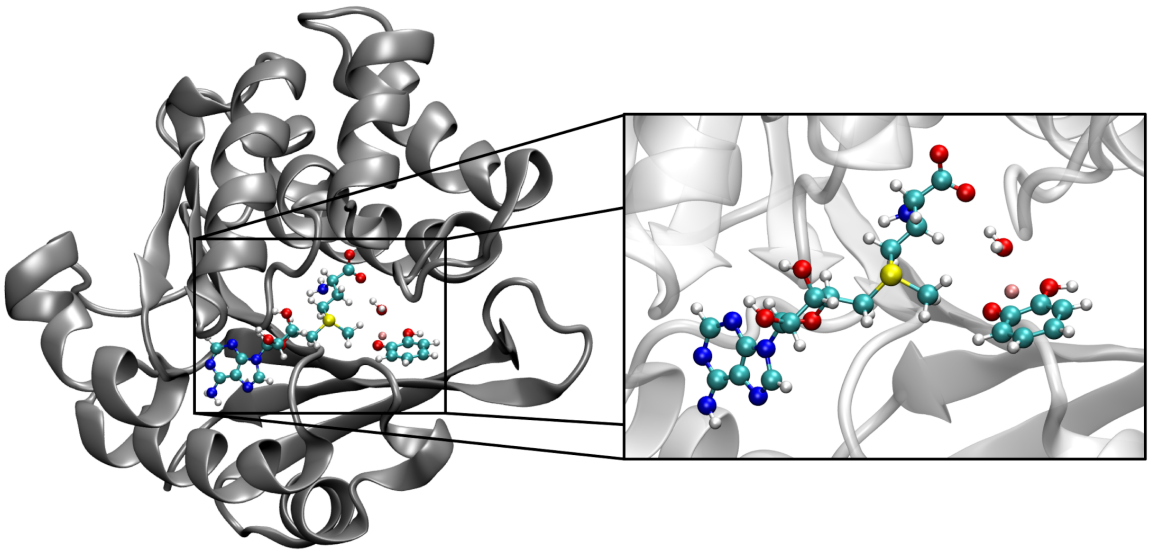


Figure 1: Visualization of catechol O-methyltransferase (COMT) with the ligands *S*-adenosyl methionine (SAM) and catecholate (CAT) as well as the catalytically active Mg^{2+} ion in the active site.

modifications were introduced concerning link atom placement and QM water, which are described in Section 2.3.

All QM calculations were performed using the GGA exchange–correlation functional PBE. In contrast to Ref. 15, we did not encounter a closing of the HUMO–LUMO gap with increasing size of the QM region. Instead, the HOMO–LUMO gap remained constant at about 1 eV for the larger QM regions (see Supporting Information, Fig. S1). Note that the results presented in Ref. 15 that will be discussed in the following, have been obtained with the range-separated hybrid functional ωPBEh, which also avoids a spurious closing of the HOMO-LUMO gap.

To test the overall QM region size convergence we first evaluated the distances between SAM methyl and the catecholate oxygen atom in the reactant structures (see Supporting Information, Fig. S2). The SAM to catechol distance in the initial MM-equilibrated structure

before QM/MM optimization is 3.11 Å which is similar to the most likely distance in the underlying distance distribution.[64] With increasing QM region size the distance significantly decreases for regions **2** and **3** from about 3.05 Å to 2.7 Å and 2.8 Å, respectively. With a distance of 2.97 Å for region **4** and **5** the SAM–CAT distance starts converging for region **6** and larger to between 2.8 and 2.9 Å. Given the differences in the QM treatment, this behavior is in reasonable agreement with the results of Ref. 15, particularly for the larger QM regions.

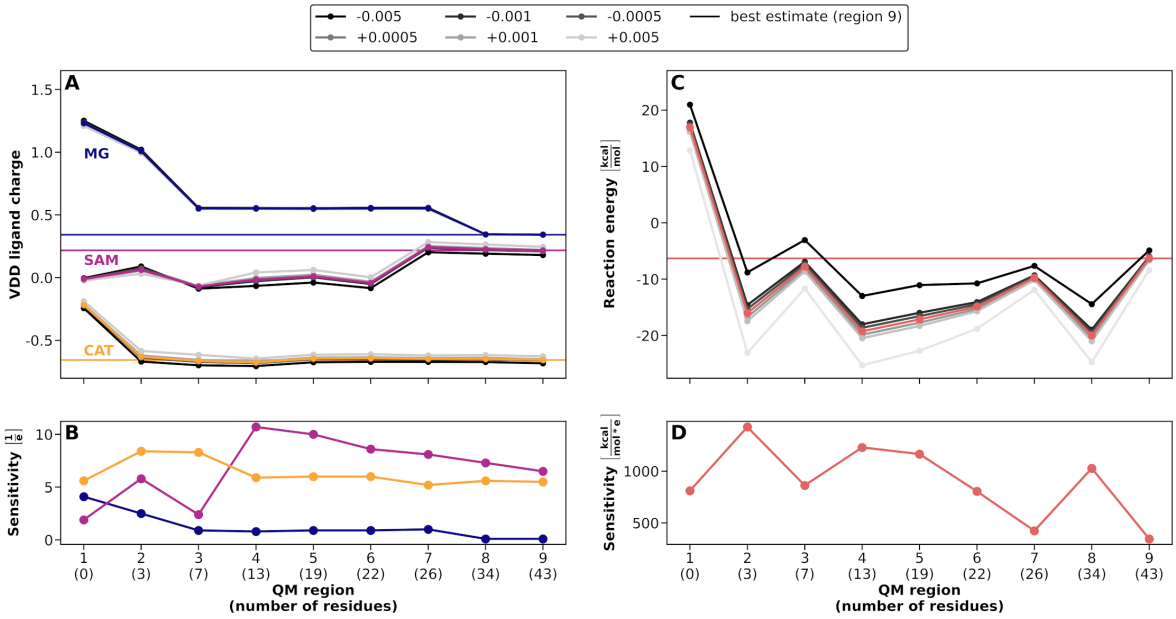


Figure 2: QM/MM convergence for QM regions constructed with the exclusively distance-based approach and corresponding global point charge variation analysis. Best estimate results (corresponding to QM region **9**) are indicated by solid horizontal lines. A: Convergence of Mg^{2+} (blue), SAM (magenta) and CAT (yellow) VDD charges with increasing QM region size. Grayscale lines indicate the ligand charges for varied MM point charges. B: Sensitivity of the VDD charges to global point charge variations $\Delta \boldsymbol{q}^{\mathrm{tot}}_{\mathrm{MM},I}$. C: Reaction energies $\Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}}$ for the methyl transfer reaction in COMT with increasing QM region size. Grayscale lines indicate the change in reaction energy with varied MM point charges. D: Sensitivity $\delta\Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}}$ of the reaction energy to global point charge variations $\Delta \boldsymbol{q}^{\mathrm{tot}}_{\mathrm{MM},I}$.

Second, we consider the Voronoi deformation density (VDD) charges of the SAM and

CAT ligands as well as the Mg^{2+} cation in the active site. Fig. 2A plots the variation of these charges with increasing QM region size. The sum of the three ligand charges is equal to +1 in the smallest QM region including only these ligands. For larger QM regions it differs from +1 because of the possibility of charge distribution over additional protein residues.

While the Mg^{2+} charge starts converging for region **3** and larger at about +0.55, the CAT and SAM charge do not converge until applying QM region **5** or larger at about −0.65 and +0.05, respectively. Again, the trends found here are in reasonable agreement with those of Ref. 15. Interestingly, the SAM charge changes to about +0.25 when going to region **7** and the Mg^{2+} charge switches to about +0.3 with region **8**. The charges obtained for region **9**, which constitute our best estimate for their converged values, are indicated in Fig. 2A as horizontal lines. For SAM, the change from a charge of around zero to +0.25 in region **7** and larger is a result of the presence of ASN91 and especially CYS94 pushing negative charge towards ILE90, which is placed directly above the adenosyl part of SAM affecting its electronic properties. The presence of ASP140, ASP168, and ASN169 completes the Mg^{2+} coordination sphere in region **8** and larger, leading to a magnesium charge of about +0.3. The absence of only one of these three residues causes the charge of +0.55.

Overall, it can be stated that small QM regions are not sufficient for reproducing the ligand charges found for large QM regions because important residues coordinating the ligands might be missing. When using a distance-based construction of the QM region, rather larger QM regions need to be reached before all relevant residues are included in the QM region.

Finally, the QM/MM reaction energy for QM regions of increasing size is shown in Fig. 2C. For small QM regions, the reaction energy shows large oscillations, while starting from QM region **4**, it steadily changes from ca. −19 kcal/mol for QM region **4** to ca. −6 kcal/mol for QM region **9**. However, this trend is broken by QM region **8**, for which the reaction energy drops to ca. −20 kcal/mol . The reaction energy for the largest QM region **9** is in reasonable agreement with the one found in Ref. 15 of ca. −10 kcal/mol and is indicated as best estimate in the figure (red horizontal line).

Overall, our results confirm the slow convergence of the reaction energy with increasing size of the QM region found in earlier studies and emphasize the need for systematic protocols for the construction and selection of the QM region. We note that Jindal and Warshel[14] found that in COMT, the activation barrier shows a much smaller sensitivity with respect to the choice of the QM region than the reaction energy. Therefore, we will not consider activation barriers and focus on the reaction energies in the present work.

4 Global Point Charge Variation Analysis for Assessing the Sensitivity of QM/MM Charges and Energies

For a first exploration of the sensitivity of the QM/MM calculations with respect to collective point-charge variations, we consider global changes of all protein MM point charges by Δq, which are compensated by an opposite change of the the solvent point charges, i.e., collective point-charge variations $\Delta \boldsymbol{q}_{\mathrm{MM}}^{\mathrm{tot}}$ as defined in Eq. (7). We will refer to this analysis as *global point charge variation analysis* (global PCVA). The effect on the QM/MM calculations is evaluated by performing single-point QM/MM calculations for the geometry optimized structures in which the MM point charges are varied. Besides the QM/MM reaction energy we consider the effect on the VDD charges of the ligands. As only single-point calculations are considered for the varied point charges, the SAM-CAT distance is excluded in the following analysis.

In addition to the VDD charges obtained with the undistorted MM charges for the different QM regions considered above, Fig. 2A includes the VDD charges obtained with collective point-charge variations $\Delta \boldsymbol{q}_{\mathrm{MM}}^{\mathrm{tot}}$ with Δq between $+0.005$ and -0.005. The corresponding sensitivities (defined in analogy to Eq. (9) for the VDD charges and evaluated numerically using a symmetric two-point finite difference formula [Eq. (13)] with $\Delta q = 0.005$) are plotted in Fig. 2B.

It can be seen that the variation of all MM point charges slightly affects the VDD charges.

Significant deviations from the unvaried curve are visible for $\Delta q = \pm 0.005$. Overall, the variations do not affect the charge convergence behavior, but lead to different ligand-dependent observations concerning the sensitivity. For CAT and Mg^{2+}, the VDD charge sensitivity decreases with increasing QM region size (with the exception of region **1** for CAT) and converges for region **4** and larger. For the largest QM regions **8** and **9**, the sensitivity of the Mg^{2+} charge is reduced to almost zero.

For SAM, in contrast to CAT and Mg^{2+}, the VDD charge sensitivity is initially increasing until a QM region size of about 300 atoms is reached (region **4**), before the sensitivity starts to slightly decrease when further enlarging the QM region. This different behavior arises because SAM is a much larger molecule than CAT and thus offers numerous possibilities for charge redistribution and more contact sites to adjacent residues. For the small QM regions, the possibilities of charge redistribution to adjacent QM residues are reduced, which results in a smaller VDD charge sensitivity in these cases. In region **4**, five residues (GLY65, TYR67, TYR70, SER71 and ILE90) which are part of the SAM coordination sphere are added to the QM region. This enables a wide variability for the SAM charge to be redistributed, resulting in an increase in sensitivity for this QM model. In larger QM regions, the SAM charge sensitivity is then gradually decreasing because only single residues being part of the SAM coordination sphere are added, such as MET39 in region **5** or TRP142 in region **6**.

The plot of the QM/MM reaction energies for the different QM regions in Fig. 2C also includes the reaction energies obtained for collective point-charge variations $\Delta \boldsymbol{q}^{\mathrm{tot}}_{\mathrm{MM}}$, whereas Fig. 2D shows the corresponding sensitivities $\delta \Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}}$. The point charge variation leads to significant changes in the reaction energy especially for a variation of $\Delta q = \pm 0.005$, for which changes of up to 8 kcal/mol could be observed for small QM regions. The convergence behavior of the energy itself is not affected by the point charge variation. The sensitivity starts decreasing strongly with region **4** while the reaction energy converges towards the one found for the largest QM region **9**. An exception is found for QM region **8**, for which the reaction energy does not follow this trend. This can be attributed to the inclusion

of four rather critical charged residues in region 8, namely the negatively charged GLU5, GLU63, and ASP168 as well as the positively charged LYS45. Remarkably, for this outlier the sensitivity is also strongly increased.

Overall, these first point charge variation tests show that small changes in the MM point charges have an impact on the QM region. While VDD charges and the reaction energy slowly converge for larger QM regions, the corresponding sensitivities to global point-charge variations decrease, whereas outliers are accompanied by an increased sensitivity. Generally, sensitivities are smaller for reaction energies closer to our best estimate (i.e., the reaction energy obtained for the largest QM region). This indicates that the sensitivity to global point-charge variations might indeed be useful as an indicator for the reliability of QM/MM calculations, and that systematically reducing this sensitivity could be a promising strategy for the systematic construction of the QM region.

5 Single Amino Acid Point Charge Variation Analysis for Systematic QM Region Construction

Motivated by the results of a global point charge variation analysis presented in the previous section, we set out to develop a protocol of the systematic construction of the QM region that aims at minimizing the sensitivity of the QM/MM reaction energy. To this end, we consider the sensitivity of the QM/MM reaction energy with respect to variations of the point charges in single amino acids, i.e., we perform a *single amino acid PCVA*. The resulting protocol for systematic QM region construction is summarized in Fig. 3 and will be described in the following.

As starting point, we consider the minimal QM/MM model with QM region **1**, i.e., for our COMT test case only the ligands and the catalytically active magnesium ion are included in the QM region. For single-point energy calculations of the geometry-optimized reactant and product structures, we calculate the sensitivity of the QM/MM energy with respect to

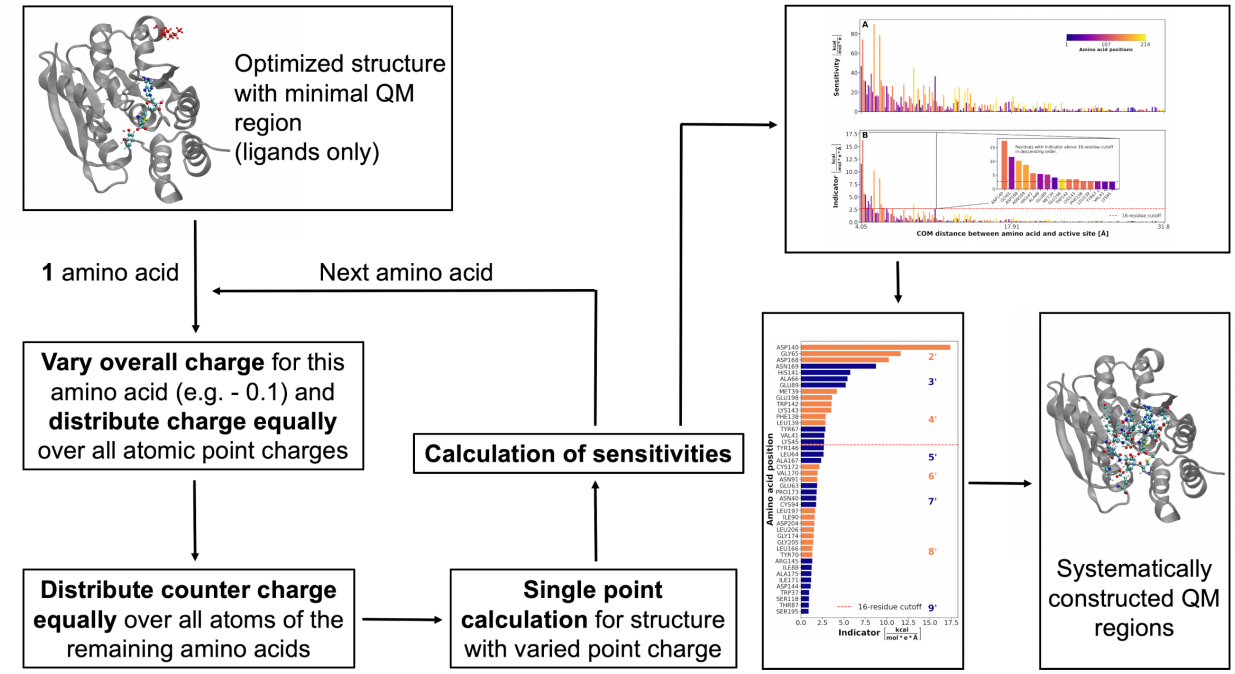


Figure 3: Schematic representation of the workflow for systematic QM region construction based on a single amino acid point charge variation analysis. See text for details.

variations of a single amino acid $\Delta q^{\mathrm{aa},i}_{\mathrm{MM},I}$ [cf. Eq. (8)]. Here, we use $\Delta q = -0.5$, i.e., the total charge of the considered amino acid is decreased by 0.5 while an equal charge of opposite sign is distributed over all other MM atoms.

To reduce the computational effort for the evaluations of this sensitivity for each amino acid, we assessed several possible simplifications compared to the global PCVA in the previous section. The different levels of simplification are referred to as PCVA-A to PCVA-E, and the sensitivities obtained in these different approximations are shown in Fig. S3. First, instead of the full QM/MM reaction energy (PCVA-A) we use only the QM contribution (PCVA-B), i.e., the approximation of Eq. (12) is employed. This is justified as the effect on the QM energy is the main focus of the analysis and possible undesired MM-only effects will be excluded. Furthermore, no significant differences in the calculated sensitivities are observed comparing the full and QM energy approach. Second, instead of the reaction energy we could use the sensitivity of the product (PCVA-C) or reactant (PCVA-D) energy instead.

While this does change the values of the sensitivities, it leads to overall similar trends (cf. Fig. S3). Using the product or reactant energy sensitivities could also be advantageous for avoiding error cancelation that might be present when considering the sensitivity of the reaction energy only. Consequently, we choose to use only the reactant energies, which reduces the number of calculations to be performed by half. The impact for this approximation on the selection of the QM region will be discussed below. Finally, instead of using a symmetric two-point formula for the numerical differentiation, it turns out to be sufficient to use a forward finite-difference formula (PCVA-E), which again reduces the number of QM calculation by another factor of two.

Altogether, we employ the PCVA-E approximation and calculate the sensitivity with respect to point-charge variations for the i-th amino acid as,

$$\delta_i E_{\mathrm{QM/MM}} = \frac{E^{\mathrm{R}}_{\mathrm{QM/MM}}\big(\boldsymbol{q}^0_{\mathrm{MM}} + \Delta\boldsymbol{q}^{aa,i}_{\mathrm{MM}}(\Delta q)\big) - E^{\mathrm{R}}_{\mathrm{QM/MM}}(\boldsymbol{q}^0_{\mathrm{MM}})}{\Delta q}, \tag{15}$$

where we employ a point charge variation of $\Delta q = -0.5$ (PCVA-E). For our COMT test case, this requires 214 QM calculations for the minimal QM region with different sets of MM point charges.

The resulting sensitivities $\delta_i E_{\mathrm{QM/MM}}$ for all single amino acid point-charge variations are shown in Fig. 4A, in which the amino acids are sorted according to their center of mass (COM) distance from the substrates in the active site of the reactant structure (QM region **1**). Naively, it could be expected that residues closer to the active site show higher sensitivities to point charge variations than distant ones. However, there are also several high-sensitivity amino acids at medium distances to the active site and low-sensitivity residues very close to the substrates. This observation confirms that an exclusively distance-based approach to include residues into the QM region is not able to detect all important amino acids and furthermore includes residues which are probably not necessary to obtain consistent QM regions.

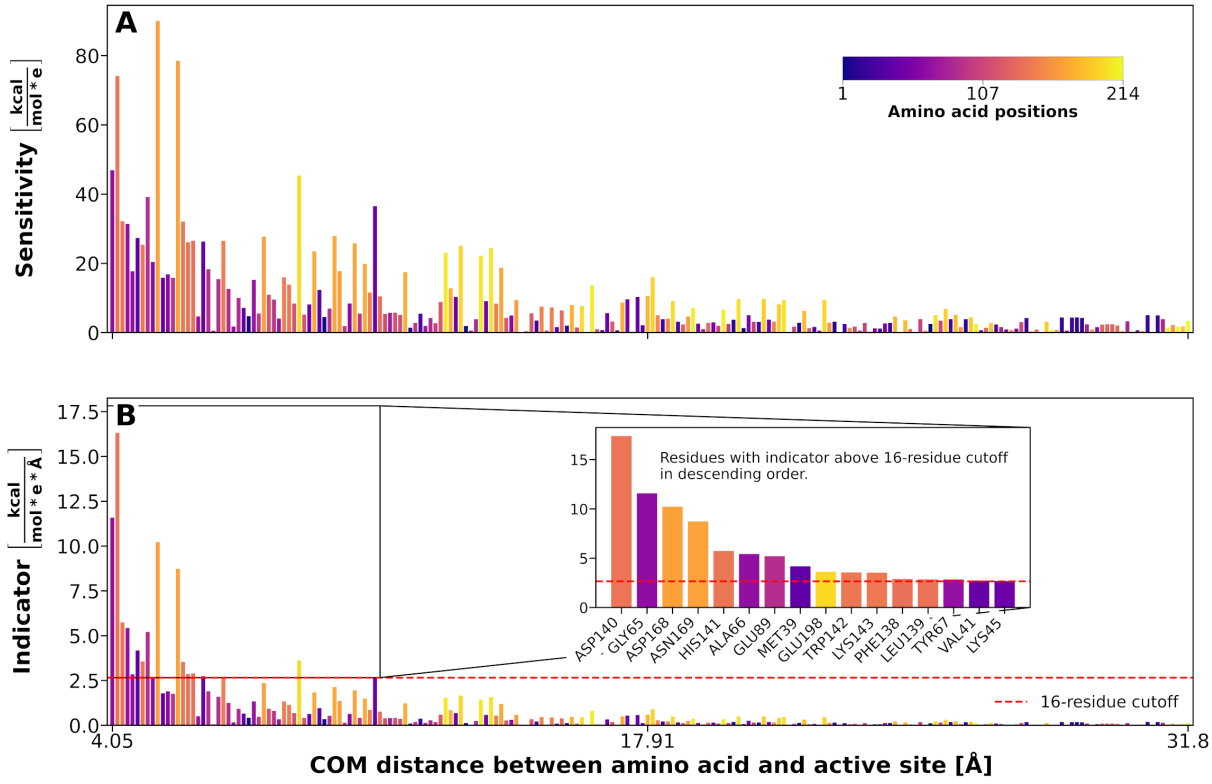


Figure 4: QM region construction based on single amino acid point charge variation analysis. A: Energy sensitivities $\delta_i E_{\mathrm{QM/MM}}$ with respect to all amino acids of COMT sorted by ascending residue–active site distance. B: Corresponding QM region indicator Θ_i for all amino acids. The 16 residues with the highest indicators are represented in the inset sorted by descending indicators.

To find a compromise between including amino acids that show a high sensitivity and those that are close to the active site, we define an empirical QM region indicator Θ_i for each amino acid by dividing the sensitivity by the COM distance between the amino acid and the active site, i.e.,

$$\Theta_i = \delta_i E_{\mathrm{QM/MM}} / \mathrm{COM}_i. \tag{16}$$

This definition ensures that distant residues with high sensitivities are considered, but also residues close to the active site which may show medium sensitivities are not overlooked. The resulting indicator is plotted in Fig. 4B. A comparison of the indicators for the schemes PCVA-A to PCVA-D is shown in the Supporting Information (Fig. S4), and an additional

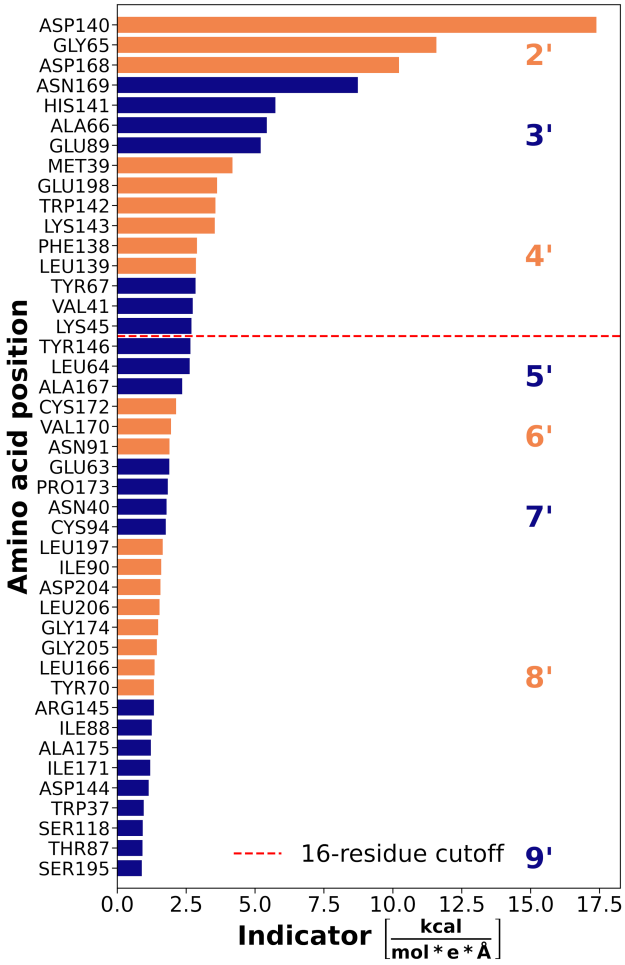


Figure 5: QM region indicators Θ_i obtained with the PCVA-E approximation sorted in descending order. Colors indicate for which QM region the corresponding residues are included. The red line indicates the 16-residue region used for comparison with CSA and FSA.

evaluation for each of these schemes analogous to Fig. 4 can be found in Figs. S5 to S8.

Table S3 lists the 16 amino acid residues with the highest QM region indicators for the schemes PCVA-A to PCVA-E (see also Section 7). When using the QM-only instead of the full QM/MM reaction energy (i.e., comparing PCVA-A and PCVA-B), the only change among the highest-ranked amino acids is that ILE88 is replaced by CYS68. Similarly, only two changes among the 16 amino acids with the highest indicators are found when using the reactant instead of the product energies (i.e., comparing PCVA-C and PCVA-D). Here, LEU64 and TRP142 are replaced by VAL41 and CYS172. All these amino acids show very

similar sensitivities and indicators in the different schemes. Larger differences are found going form the reaction energies (PCVA-B) to using the reactant energies (PCVA-D), where six differences appear among the top-16 amino acids. However, these changes mostly occur for amino acids with very close values of the QM region indicator. Furthermore, the use of the reactant or product energies avoids error cancelation that might be present when using the reaction energy. Finally, only two differences (VAL41 and TYR67 instead of LEU64 and TYR146) are found among the 16 highest-ranked amino acids when going to a simplified numerical differentiation formula (i.e., from PCVA-D to PCVA-E). Overall, this confirms that the PCVA-E approximation seems to be a reasonable choice.

By including the amino acids with the highest QM region indicators Θ_i (see Fig. 5), we are now able to systematically construct QM regions that should reduce the sensitivities of the QM/MM reaction energy to point charge variations.

6 Assessment of PCVA-Based QM Regions of Increasing Size

As a first test, we assess the PCVA-based construction of QM regions with increasing size. Ideally, by using PCVA for a systematic construction of QM regions, the convergence towards the results obtained with very large QM regions should be accelerated compared to an exclusively distance-based construction of the QM region. To this end, we construct QM regions consisting of just as many residues as in the distance-based approach (see Section 4) and label these regions as **2'**, **3'** etc. (e.g., QM regions **3** and **3'** both contain seven amino acid residues). Fig. 5 and Tab. S4 in the Supporting Information show which residues are included for each QM region. Again, geometry optimizations of the reactant and product structures were performed for each of these QM regions.

Fig. 6 shows the convergence of the VDD charges and the QM/MM reaction energy as well as the corresponding sensitivities to global point charge variations (cf. Fig. 2 for the same plot

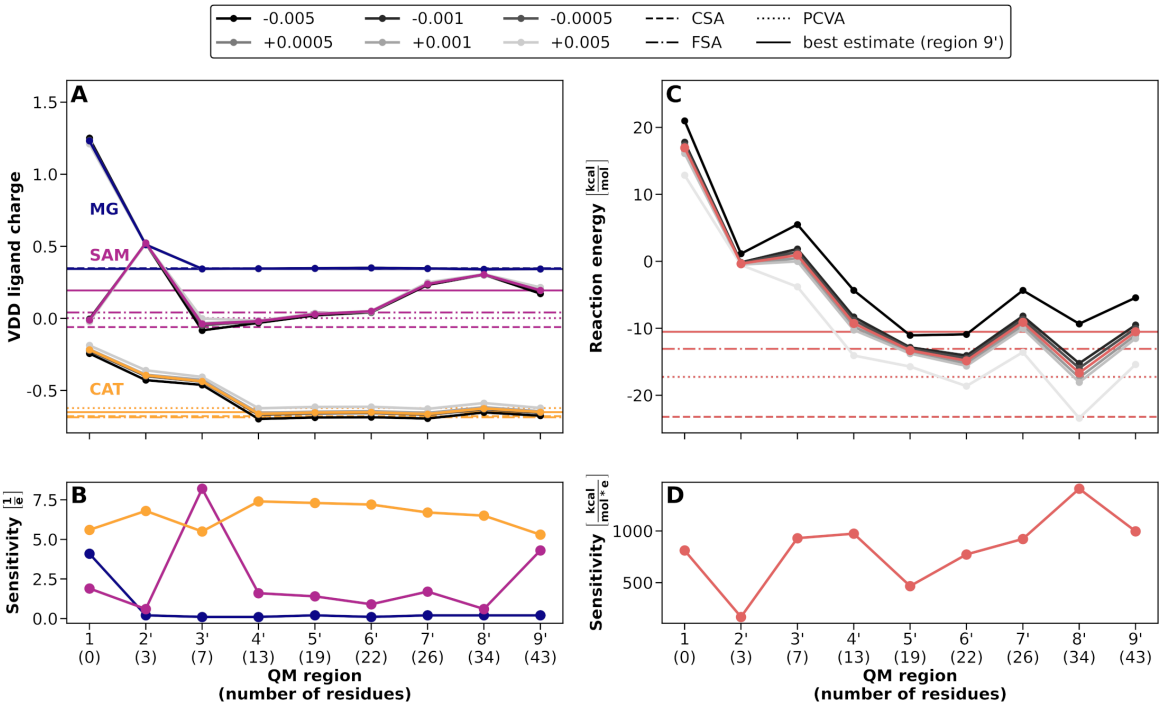


Figure 6: QM/MM convergence for QM regions constructed based on PCVA and corresponding global point charge variation analysis. Best estimate results (corresponding to QM region **9**') are indicated by solid horizontal lines. Results for atom-economical QM regions consisting of 16 amino acids based on PCVA, CSA, and FSA are included as dotted, dashed, and dashdotted horizontal lines, respectively. A: Convergence of Mg^{2+} (blue), SAM (magenta) and CAT (yellow) VDD charges with increasing QM region size. Grayscale lines indicate the ligand charges for varied MM point charges. B: Sensitivity of the VDD charges to global point charge variations $\Delta \boldsymbol{q}^{\mathrm{tot}}_{\mathrm{MM},I}$. C: Reaction energies $\Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}}$ for the methyl transfer reaction in COMT with increasing QM region size. Grayscale lines indicate the change in reaction energy with varied MM point charges and the horizontal line indicates our best estimate. D: Sensitivity $\delta\Delta E^{\mathrm{reaction}}_{\mathrm{QM/MM}}$ of the reaction energy to global point charge variations $\Delta \boldsymbol{q}^{\mathrm{tot}}_{\mathrm{MM},I}$.

for the distance-based construction of QM regions). The VDD charges (see Fig. 6A) converge to similar values as for the distance-based construction of the QM region. Remarkably, the Mg^{2+} charge converges to about +0.3 already for region **3'** and larger, which was achieved for the distance-based inclusion of residues only with region **8**. The SAM charges behave

similarly to the distance-based case. From region **3'** onwards, it stabilizes at around zero, before reaching +0.25 in region **7'** and larger. The CAT charges again starts converging for region **4'** and larger.

PCVA-constructed QM regions deliver overall lower sensitivities regarding VDD ligand charges (see Fig. 6B), which marks an expected behavior because high-sensitivity residues are included in the QM region by PCVA. The convergence of the charge sensitivities is similar to the analysis for distance-based QM regions (cf. Fig. 2) with a fast convergence for Mg^{2+}, a constant behavior for CAT, and jumps for SAM with a slightly decreasing trend.

The reaction energy (see Fig. 6C) decreases for the smallest QM regions, starts to stabilize for QM regions **4'**, and oscillates around our best estimate of about −11 kcal/mol for larger QM regions. However, these oscillations are smaller than for the distance-based construction of the QM regions. Note that compared to Fig. 2, we updated our best estimate to correspond to QM region **9'** in Fig. 6C.

In contrast to the distance-based case, a slightly increasing trend is observed for the reaction energy sensitivity (Fig. 6D). For QM region **8'**, at which the reaction energy is further from our best estimate, the sensitivity also shows a marked increase. Except for the smallest QM regions, the global PCVA sensitivity can thus be used to judge the accuracy of the calculated QM/MM reaction energy. The overall larger sensitivity for the PCVA-based QM regions compared to the distance-based construction could be an effect of a larger contact area between the MM and QM regions because in the PCVA case, MM residues close to the active site with low sensitivities remain in the MM part and can affect adjacent QM residues.

7 Assessment of Atom-Economical QM Region Based on PCVA

We now consider the determination of an atom-economical QM region consisting of 16 amino acids for COMT based on PCVA. This approach follows the work of Kulik *et al.*, who constructed such QM regions using their CSA and FSA approach,[15,19] and thus allows for a direct comparison to these approaches. To this end, we included the 16 amino acids with the highest QM region indicators Θ_i in the QM region (see Fig. 5), and again performed QM/MM geometry optimizations of the reactant and product structures. Similarly, we performed QM/MM geometry optimizations using the QM regions obtained in Refs. 15,19 which also each include 16 amino acids.

Table 1 compares the amino acid residues included in such an atom-economical QM region for a distance-based QM region construction, the CSA and FSA approaches, and for our PCVA-based QM-region construction.

Compared to the CSA and FSA approaches applied by Kulik *et al.*[15,19] our PCVA approach detects the majority of the residues that should be part of the 16-residue QM region, as it can be seen in Table 1 (e.g., VAL41, GLU89, or ASP140). Moreover, also more distant residues are included that would not be considered in a 16-residue QM region in the distance-based case (e.g., MET39, LYS45, ALA66, or ASP168). This indicates the ability of PCVA to even detect high-impact residues located relatively distant from the active site COM. Nevertheless, there are several residues included by CSA and/or FSA that are not under the 16 highest-ranked residues when applying the PCVA approach. However, most of these residues are assigned a PCVA rank close to 16 such as ASN40, GLU63, or ILE90.

An extreme case with a very low PCVA rank (178) compared to CSA (10) and FSA (10-11) is SER71. This residue is located very close to the active site and thus potentially important for ligand binding, but it shows a very low electrostatic effect on the substrates and is thus not detected by PCVA. Another case is SER118, which is ranked about 30 places

Table 1: Comparison of residues included in an atom-economical 16-residue QM region for a distance-based approach, CSA, FSA and the PCVA-E approach. Numbers indicate the rank of the amino acid according the value assigned by the corresponding scheme. For residues with the same value or if the values are not given in the cited references, a range of ranks is given. The highest-ranked 16 residues for CSA, FSA and PCVA are marked in red. An extended version of this table including PCVA-A to PCVA-D can be found in the Supporting Information (Tab. S2). Table S3 gives lists the compositions of the different 16-residue QM regions.

Residue	**Distance-based**[15]	**CSA**[15]	**FSA**[19]	**PCVA**
MET39	14-19	14-16	14	8
ASN40	23-26	12-13	7	25
VAL41	1-3	8	4	15
LYS45	27-34	17	18-155	16
GLU63	27-34	3-4	18-155	23
GLY65	8-13	24-36	2	2
ALA66	20-22	6-7	1	6
TYR67	8-13	14-16	10-11	14
TYR70	8-13	24-36	5	34
SER71	8-13	10	10-11	178
ALA72	35-43	9	18-155	86
GLU89	4-7	1	13	7
ILE90	8-13	14-16	12	28
SER118	4-7	12-13	8	41
PHE138	14-19	24-36	156-163	12
LEU139	44-56	24-36	156-163	13
ASP140	4-7	2	3	1
HIS141	8-13	11	9	5
TRP142	20-22	37+	17	10
LYS143	1-3	37+	18-155	11
ASP168	27-34	6-7	16	3
ASN169	4-7	3-4	6	4
GLU198	1-3	5	15	9

lower in PCVA than in CSA and FSA. Both these cases concern serine residues, which might show only a small electrostatic effect.

PCVA also detects residues that are much lower ranked in CSA and FSA. PHE138 and LEU139 are both part of the ligand binding site and obviously have a high electrostatic impact on the substrates. Consequently, PCVA ranks them very high at 12 and 13, respectively, in contrast to FSA (156-163) while CSA also assigns quite high ranks to these residues (24-36). Similar results are observed for TRP142 and LYS143, which also seem to play a role in ligand binding and are ranked on 10 and 11 in PCVA, whereas FSA assigns much lower ranks (17 and 18-155) than CSA (37+).

The shown differences between PCVA and CSA or FSA do not indicate that PCVA is performing worse as none of the existent methods can be considered the gold standard. Furthermore, even between CSA and FSA major differences in the evaluation of residues can be observed (e.g. for GLU63, GLY65, TYR70, or ALA72) leading to differently composed QM regions.

In Fig. 6, we compare the VDD charges of SAM, CAT, and the catalytically-active Mg^{2+} cation as well as the QM/MM reaction energy obtained for the atom-economical QM regions constructed using CSA, FSA, and PCVA to those obtained for PCVA-constructed QM regions of increasing size discussed in Section 6. Note that the size of the atom-economical QM regions of 16 amino acids is in between those of QM regions **4'** (13 amino acids) and **5'** (19 amino acids).

Regarding VDD charges (see Fig. 6A), PCVA performs as well as CSA and FSA. The change in the Mg^{2+} charge to about 0.3 with the distance-based region **8** (34 residues) and larger is achieved already in the QM regions containing 16 residues. For SAM, all three atom-economical QM regions also provide the converged charge. In the SAM case, none of the methods is able to detect the charge change to 0.25 for region **7** and larger. This is a fundamental limitation of QM regions of a given size and can probably only be rectified by using larger QM regions.

Concerning the reaction energy (Fig. 6B), PCVA and FSA deliver reasonable values compared to the large-region results (best estimate based on region **9'** of −10.5 kcal/mol) with about −17.5 kcal/mol for PCVA and −13 kcal/mol for FSA. CSA performs worse with a reaction energy of about −23 kcal/mol. Overall, an atom-economical 16-residue QM region constructed based on PCVA delivers reasonable results close to our best estimate based on the largest QM region.

8 Conclusion

As a first step towards systematically quantifying the uncertainties of QM/MM calculations, we have presented a point charge variation analysis for assessing the sensitivity of QM/MM reaction energies to changes of the MM point charges. To this end, we considered the derivative of the QM/MM reaction energy with respect to selected distortions of the MM charges. This derivative can be calculated numerically by performing QM/MM calculations with varied MM point charges, and different efficient approximations can be employed. Generally, the most simple approximation (PCVA-E), in which a forward finite difference is used and only the reactant structure is considered, turns out to be sufficient for a qualitative assessment of uncertainties.

A global PCVA, in which all protein point charges are varied simultaneously, can be used as a simple indicator of the accuracy of the resulting QM/MM reaction energy as well as other properties of the active site, such as ligand charges. For the considered test cases we found that when comparing different QM regions, the sensitivity in a global PCVA is smaller for those QM regions that yield results closer to the best estimate obtained for the largest QM regions, i.e., the sensitivity generally decreases when approaching the limit of a full QM calculation. However, this only holds once the QM regions have reached a reasonable size and the correlation between the sensitivity in a global PCVA and the deviation from the best estimate is not always clear.

Nevertheless, our results demonstrate that the analysis of the sensitivity with respect to the MM point charges is a good starting point for the investigation of uncertainties in QM/MM calculations. We are planning to extent this work in the future by considering not only selected distortions of the MM point charges, but performing a full sensitivity analysis that allows one to identify the collective point charge variations that have the largest influence on QM/MM reaction energies, following our earlier work on structural distortions in theoretical spectroscopy.[40] While the reliability of considering only one global distortion of the MM point charges is limited, such a more comprehensive assessment of the sensitivity can be expected to overcome the limitations of our present approach.

In addition to a global PCVA, we have considered a single amino acid PCVA, in which the MM charges of each amino acid residue are varied. This makes it possible to assess the contribution of each amino acid to the uncertainty in the QM/MM reaction energy, and can be used to guide the systematic construction of the QM region. By including amino acids with a high sensitivity to point charge variations in the QM region, the overall sensitivity of the QM/MM reaction energy can be reduced. Here, we devised a PCVA-based scheme for the systematic construction of the QM region.

For the considered test case, our scheme leads to a faster and more reliable convergence with the size of the QM region compared to distance-based QM region construction. Comparing to atom-economical QM regions of the same size provided by the other common approaches, in particular CSA and FSA, our PCVA-based approach performed well and yields similar QM regions.

The huge advantage of PCVA is its much lower computational cost compared to the CSA or FSA approach (see Supporting Information, Tab. S5). Our PCVA-based approach requires only a geometry optimization of the target system with a minimal QM region including substrates, which is followed by single-point calculations for the point-charge variation of each amino acid. In contrast, CSA is a very expensive approach based on large QM regions with up to 1000 atoms. For these large systems, geometry optimizations have to be performed

for the holo and apo enzyme structure for several snapshots along the reaction coordinate.[15] FSA, even though being much cheaper than CSA, still needs as many geometry optimization as there are amino acids in the system for the minimal QM region plus one additional residue in each calculation.[19] A rather similar approach to our PCVA-based construction of the QM region, the charge deletion analysis (CDA), is mostly reported for the usage with medium-sized QM regions,[18,65–70] which also increases the required computational effort. Of course, our PCVA-based approach can also be applied for larger QM regions, but we found that this does not improve the results substantially.

The PCVA-based construction of the QM region is limited to the electrostatic effect of the amino acids and thereby lacking other properties which may also play an important role in QM region determination. Therefore, it is possible that crucial residues (e.g. catalytically important) may be absent under the detected residues. Here, the biochemical and structural understanding should be considered as well when constructing QM regions. Altogether, we suppose that our fast and computationally cheap approach is a good complement to existing methods for the automatic and systematic QM region construction. We expect that future developments concerning uncertainty quantification for QM/MM calculation will also allow for the development of more sophisticated schemes for the systematic construction of QM regions.

Supporting Information

Additional tables with details about the composition of distance-based, PCVA-based and 16-residue QM regions, an extended version of Table 1, and a comparison of the computational effort. Additional figures with information about the HOMO-LUMO gap, the SAM-CAT distance convergence, and the comparison of sensitivities and indicators for the different approximate PCVA schemes.

References

(1) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(2) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **1990**, *11*, 700–733.

(3) Gao, J. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995; Vol. 7; pp 119–185.

(4) Groenhof, G. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2013; pp 43–66, DOI: 10.1007/978-1-62703-017-5_3.

(5) Friesner, R. A.; Guallar, V. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

(6) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(7) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Computational tools for the evaluation of laboratory-engineered biocatalysts. *Chem. Commun.* **2016**, *53*, 284–297.

(8) Ryde, U. In *Methods in Enzymology*; Voth, G. A., Ed.; Computational Approaches for Studying Enzyme Mechanism Part A; Academic Press, 2016; Vol. 577; pp 119–158.

(9) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. a. Application of quantum mechanics/molecular

mechanics methods in the study of enzymatic reaction mechanisms. *WIREs Comput. Mol. Sci.* **2017**, *7*, e1281.

(10) Ahmadi, S.; Herrera, L. B.; Chehelamirani, M.; Hostaš, J.; Jalife, S.; Salahub, D. R. Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *Int. J. Quantum Chem.* **2018**, *118*, e25558.

(11) Magalhães, R. P.; Fernandes, H. S.; Sousa, S. F. Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges. *Isr. J. Chem.* **2020**, *60*, 655–666.

(12) Sumowski, C. V.; Ochsenfeld, C. A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113*, 11734–11741.

(13) Flaig, D.; Beer, M.; Ochsenfeld, C. Convergence of Electronic Structure with the Size of the QM Region: Example of QM/MM NMR Shieldings. *J. Chem. Theory Comput.* **2012**, *8*, 2260–2271.

(14) Jindal, G.; Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **2016**, *120*, 9913–9921.

(15) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381–11394.

(16) Mehmood, R.; Kulik, H. J. Both Configuration and QM Region Size Matter: Zinc Stability in QM/MM Models of DNA Methyltransferase. *J. Chem. Theory Comput.* **2020**, *16*, 3121–3134.

(17) Sumner, S.; Söderhjelm, P.; Ryde, U. Effect of Geometry Optimizations on QM-Cluster

and QM/MM Studies of Reaction Energies in Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4205–4214.

(18) Liao, R.-Z.; Thiel, W. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *J. Comput. Chem.* **2013**, *34*, 2389–2397.

(19) Karelina, M.; Kulik, H. J. Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **2017**, *13*, 563–576.

(20) Brunken, C.; Reiher, M. Automated Construction of Quantum–Classical Hybrid Models. *J. Chem. Theory Comput.* **2021**, *17*, 3797–3813.

(21) Roßbach, S.; Ochsenfeld, C. Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *J. Chem. Theory Comput.* **2017**, *13*, 1102–1107.

(22) Cao, L.; Ryde, U. On the Difference Between Additive and Subtractive QM/MM Calculations. *Front. Chem.* **2018**, *6*, 89.

(23) Solt, I.; Kulhánek, P.; Simon, I.; Winfield, S.; Payne, M. C.; Csányi, G.; Fuxreiter, M. Evaluating Boundary Dependent Errors in QM/MM Simulations. *J. Phys. Chem. B* **2009**, *113*, 5728–5735.

(24) Loco, D.; Lagardère, L.; Adjoua, O.; Piquemal, J.-P. Atomistic Polarizable Embeddings: Energy, Dynamics, Spectroscopy, and Reactivity. *Acc. Chem. Res.* **2021**, *54*, 2812–2822.

(25) Nochebuena, J.; Naseem-Khan, S.; Cisneros, G. A. Development and application of quantum mechanics/molecular mechanics methods with advanced polarizable potentials. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1515.

(26) Zhang, Y.; Lin, H. Flexible-Boundary Quantum-Mechanical/Molecular-Mechanical

Calculations: Partial Charge Transfer between the Quantum-Mechanical and Molecular-Mechanical Subsystems. *J. Chem. Theory Comput.* **2008**, *4*, 414–425.

(27) Zhang, Y.; Lin, H. Flexible-boundary QM/MM calculations: II. Partial charge transfer across the QM/MM boundary that passes through a covalent bond. *Theor. Chem. Acc.* **2010**, *126*, 315–322.

(28) Maseras, F.; Morokuma, K. IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(29) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and Pt(P(t-Bu)$_3$)$_2$ + H_2 Oxidative Addition. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(30) Swart, M. AddRemove: A new link model for use in QM/MM studies. *Int. J. Quantum Chem.* **2003**, *91*, 177–183.

(31) Philipp, D. M.; Friesner, R. A. Mixed ab initio QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide. *J. Comput. Chem.* **1999**, *20*, 1468–1494.

(32) Assfeld, X.; Rivail, J.-L. Quantum chemical computations on parts of large molecules: the ab initio local self consistent field method. *Chem. Phys. Lett.* **1996**, *263*, 100–106.

(33) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations. *J. Phys. Chem. A* **1998**, *102*, 4714–4721.

(34) Smith, R. *Uncertainty Quantification: Theory, Implementation, and Applications*; Society for Industrial and Applied Mathematics: Philadelphia, 2014.

(35) Sullivan, T. J. *Introduction to Uncertainty Quantification*, 1st ed.; Springer: New York, NY, 2015.

(36) Irikura, K. K.; Johnson III, R. D.; Kacker, R. N. Uncertainty associated with virtual measurements from computational quantum chemistry models. *Metrologia* **2004**, *41*, 369.

(37) Glotzer, S. C.; Kim, S.; Cummings, P. T.; Deshmukh, A.; Head-Gordon, M.; Karniadakis, G.; Petzold, L.; Sagui, C.; Shinozuka, M. WTEC Panel Report on International Assessment of Research and Development in Simulation-Based Engineering and Science. 2013; DOI: 10.2172/1088842, URL: http://www.osti.gov/servlets/purl/1088842/.

(38) Simm, G. N.; Proppe, J.; Reiher, M. Error Assessment of Computational Models in Chemistry. *Chimia* **2017**, *71*, 202–208.

(39) Oung, S. W.; Rudolph, J.; Jacob, Ch. R. Uncertainty quantification in theoretical spectroscopy: The structural sensitivity of X-ray emission spectra. *Int. J. Quantum Chem.* **2018**, *118*, e25458.

(40) Bergmann, T. G.; Welzel, M. O.; Jacob, Ch. R. Towards theoretical spectroscopy with error bars: systematic quantification of the structural sensitivity of calculated spectra. *Chem. Sci.* **2020**, *11*, 1862–1877.

(41) Cacuci, D. G. *Sensitivity & Uncertainty Analysis, Volume 1: Theory*, 1st ed.; Chapman and Hall/CRC: Boca Raton, 2003.

(42) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(43) GROMACS Version 2019.6. 2020; DOI: 10.5281/zenodo.3685922, URL: http://www.gromacs.org/.

(44) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(46) Software for Chemistry and Materials, Amsterdam, Ams, Amsterdam Modelling Suite. 2020; URL: http://www.scm.com.

(47) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931–967.

(48) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(49) Van Lenthe, E.; Baerends, E. J. Optimized Slater-type basis sets for the elements 1-118. *J. Comput. Chem.* **2003**, *24*, 1142–1156.

(50) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(51) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(52) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Model.* **2006**, *25*, 247–260.

(53) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012**, *5*, 367.

(54) Acpype. 2021; URL: https://github.com/alanwilter/acpype.

(55) Software for Chemistry and Materials, Hybrid Engine Manual — Hybrid 2020 documentation. 2020; URL: https://www.scm.com/doc.2020/Hybrid/index.html.

(56) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbsch, P. Structural Relaxation Made Simple. *Phys. Rev. Lett.* **2006**, *97*, 170201.

(57) Fonseca Guerra, C.; Handgraaf, J.-W.; Baerends, E. J.; Bickelhaupt, F. M. Voronoi deformation density (VDD) charges: Assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD methods for charge analysis. *J. Comput. Chem.* **2004**, *25*, 189–210.

(58) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(59) Matplotlib Version 3.4.2. 2021; DOI: 10.5281/zenodo.4743323, URL: https://matplotlib.org.

(60) Humphrey, W.; Dalke, A.; Schulten, K. VMD — Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(61) Brandt, F.; Jacob, Ch. R. Data Set: Systematic QM Region Construction in QM/MM Calculations Based on Uncertainty Quantification. 2021; DOI: 10.5281/zenodo.5618058.

(62) Axelrod, J.; Tomchick, R. Enzymatic O-Methylation of Epinephrine and Other Catechols. *J. Biol. Chem.* **1958**, *233*, 702–705.

(63) Rutherford, K.; Le Trong, I.; Stenkamp, R. E.; Parson, W. W. Crystal Structures of Human 108V and 108M Catechol O-Methyltransferase. *J. Mol. Biol.* **2008**, *380*, 120–130.

(64) Zhang, J.; Kulik, H. J.; Martinez, T. J.; Klinman, J. P. Mediation of donor–acceptor distance in an enzymatic methyl transfer reaction. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 7954–7959.

(65) Bash, P. A.; Field, M. J.; Davenport, R. C.; Petsko, G. A.; Ringe, D.; Karplus, M. Computer simulation and analysis of the reaction pathway of triosephosphate isomerase. *Biochemistry* **1991**, *30*, 5826–5832.

(66) Mulholland, A. J.; Richards, W. G. Acetyl-CoA enolization in citrate synthase: A quantum mechanical/molecular mechanical (QM/MM) study. *Proteins: Struct., Funct., Bioinf.* **1997**, *27*, 9–25.

(67) Mladenovic, M.; Fink, R. F.; Thiel, W.; Schirmeister, T.; Engels, B. On the Origin of the Stabilization of the Zwitterionic Resting State of Cysteine Proteases: A Theoretical Study. *J. Am. Chem. Soc.* **2008**, *130*, 8696–8705.

(68) Gómez, H.; Polyak, I.; Thiel, W.; Lluch, J. M.; Masgrau, L. Retaining Glycosyltransferase Mechanism Studied by QM/MM Methods: Lipopolysaccharyl-?-1,4-galactosyltransferase C Transfers ?-Galactose via an Oxocarbenium Ion-like Transition State. *J. Am. Chem. Soc.* **2012**, *134*, 4743–4752.

(69) Liao, R.-Z.; Thiel, W. Why Is the Oxidation State of Iron Crucial for the Activity of Heme-Dependent Aldoxime Dehydratase? A QM/MM Study. *J. Phys. Chem. B* **2012**, *116*, 9396–9408.

(70) Liao, R.-Z.; Thiel, W. Comparison of QM-Only and QM/MM Models for the Mechanism of Tungsten-Dependent Acetylene Hydratase. *J. Chem. Theory Comput.* **2012**, *8*, 3793–3803.