

UC Davis

UC Davis Previously Published Works

Title

A Polarizable QM/MM Model That Combines the State-Averaged CASSCF and AMOEBA Force Field for Photoreactions in Proteins

Permalink

<https://escholarship.org/uc/item/0766r05r>

Authors

Song, Chenchen

Wang, Lee-Ping

Publication Date

2024-08-01

DOI

10.1021/acs.jctc.4c00623

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

A polarizable QM/MM model that combines state-averaged CASSCF and AMOEBA force field for photoreactions in proteins

Chenchen Song* and Lee-Ping Wang

Department of Chemistry, University of California, Davis; 1 Shields Ave; Davis, CA 95616.

E-mail: ccsong@ucdavis.edu

Abstract

This paper presents the polarizable quantum mechanics/molecular mechanics (QM/MM) embedding of state-averaged complete active space self-consistent field (SA-CASSCF) in the atomic multipole optimized energetics for biomolecular applications (AMOEBA) force field for the purpose of studying photoreactions in protein environments. We describe two extensions of our previous work that combines SA-CASSCF with AMOEBA water models allowing it to be generalized to AMOEBA models for proteins and other macromolecules. First, we discuss how our QM/MM model accounts for the discrepancy between the direct and polarization electric field that arises in the AMOEBA description of intramolecular polarization. A second improvement is the incorporation of link-atom schemes to treat instances where the QM/MM boundary goes through covalent bonds. A single link atom scheme and double link atom scheme are considered in this work, and we will discuss how electrostatic interaction, van der Waals interaction and various kinds of valence terms are treated across the boundary. To test the accuracy of the link atom scheme, we will compare QM/MM with full QM

calculations, and study how the errors in ground state properties, excited state properties, and excitation energies change when tuning the parameters in the link atom scheme. We will also test the new SA-CASSCF/AMOEBA method on an elementary reaction step in NanoLuc, an artificial bioluminescence luciferase. We will show how the reaction mechanism is different when calculated in the gas phase, in polarizable continuum medium (PCM), versus in protein AMOEBA models.

1 Introduction

Light plays a major role in chemistry and biochemistry, both as a source of energy to cross over thermally inaccessible energy barriers, and also an important means for organisms to sense their environments and exchange signals. Plants and animals have evolved to harness UV-visible light in biological functions,¹ and at the same time have also evolved mechanisms to mitigate the detrimental effects from uncontrolled photochemistry.² A major subset of photoreactions in life relies on proteins that contain a small photoactive molecule (the chromophore), such as the bilin in phytochrome for plant red light regulation,³ retinal in rhodopsin for vision,⁴ and the coelentrazine in luciferase for bioluminescence.⁵ The molecules and strategies employed by Nature to harness light are also used to inspire artificial light-harvesting and light-activated technologies.⁶

Due to the ultrafast timescale of the photoreactions and the small size of chromophores relative to the proteins, experimental investigation of photoreactions in biomolecules often requires high spatial and temporal resolution. Recently, it has become possible to obtain experimental data via techniques such as time-resolved serial femtosecond X-ray crystallography,^{7,8} but these experiments are technically challenging and require advanced infrastructure such as X-ray free electron lasers. Therefore, theoretical methods, especially ab initio quantum chemistry, have the potential to provide important insights into the mechanisms of these reactions within the macromolecular environment.

Because the electronic properties change qualitatively upon excitations, quantum me-

chanics is required for studying photoreactions. However, the sheer size of these photoactive proteins makes it infeasible to describe the entire system quantum mechanically. Fortunately, because the excited wavefunction is primarily localized in the chromophore region, one can easily adopt a QM/MM hybrid model^{9,10} that treats a small portion of the system (i.e. the chromophore and its immediate surroundings) with quantum mechanics (QM), and the remaining environment using classical molecular mechanics (MM). The environment is important because electrostatic and polarization effects can modulate the relative energies of ground and excited states, and these effects along with steric effects can change the shape of the potential energy surfaces. Such hybrid QM/MM models make it possible to apply advanced QM methods to study photoreactions in complex environments such as proteins.

When it comes to choose the level of theory, the advantages of using multi-reference electronic structure methods to study photoreactions have already been extensively reviewed.^{11,12} Multi-reference methods can cover a wide range of potential energy surfaces (PES) especially in regions with prominent static correlation effects (e.g. bond breaking) and can provide the correct topology of conical intersections,¹³ making them the ideal choice of the QM method. In this work, we use the state-averaged complete active space self consistent field (SA-CASSCF) method.¹⁴⁻¹⁶ SA-CASSCF can often produce qualitatively correct PES at relatively low computational cost, thus has been widely used in non-adiabatic molecular dynamics simulations.^{17,18} In addition, SA-CASSCF is also a necessary step to provide the reference wavefunctions needed for even higher level of theories such as multi-reference perturbation theory¹⁹ and multi-reference configuration interaction.²⁰ In terms of the classical models for the protein environment, many previous studies have pointed out that the polarizable force fields stand out in their abilities to capture a wide range of environmental effects important for light-driven processes.^{21,22} In particular, polarizable force fields combined with molecular dynamics (MD) can capture the nonequilibrium polarization effects, where the induced charges or dipoles mimic the fast reponse from the environment electrons, while the structural changes along the MD trajectory mimic the time-dependent slow response

from the nuclear motions. In addition, since polarizable force fields preserve the molecular structures, environment effects such as hydrogen bonding, viscosity, and steric effects can also be described.^{23,24} In this work, we choose the atomic multipole optimized energetics for biomolecular applications (AMOEBA)^{25,26} polarizable force field for describing the protein environment. The AMOEBA force field is well suited to QM/MM simulations due to its detailed electrostatic model that can fit QM electrostatic potentials to high accuracy. AMOEBA also has a development history of over 20 years which has produced parameters available for proteins,²⁷ nucleic acids,²⁸ and small molecules,²⁹ new methods and technologies for improving simulation performance,³⁰ automated parameterization workflows,³¹ and adaptations to QM/MM simulations especially in the context of single reference quantum chemistry method such as density functional theory (DFT)^{32,33} and time-dependent density functional theory (TD-DFT).^{34,35}

In our previous work, we have successfully developed polarizable embedding of SA-CASSCF in AMOEBA water models to simulate photoreactions in solvents.^{36,37} By formulating the polarization energy of the AMOEBA water model as a variational problem,³⁸ we formulated self-consistent field equations for SA-CASSCF/AMOEBA, i.e. the wavefunction parameters and induced dipoles are solved by variationally minimizing the state-averaged energy, which is a natural generalization of the standard SA-CASSCF framework. In addition, we have incorporated the dynamic weight scheme³⁹ into the definition of the state averaged energy, which provides smooth interpolation between a state-specific description important for modeling state-specific polarization in the Franck-Condon region and an equally state-averaged description at the conical intersection region ensuring its correct topology. We have also developed the corresponding nuclear gradients and non-adiabatic couplings, and have interfaced it with ab initio multiple spawning^{40,41} to carry out nonadiabatic molecular dynamics simulations.

In order to generalize SA-CASSCF/AMOEBA from solvents to protein environments, there are two challenges that need to be addressed. One challenge is that the AMOEBA

model can no longer be formulated as a variational approach when intramolecular polarization is included, and such effect cannot be neglected for macromolecules. The breakdown of the variational formulation is due to a discrepancy within AMOEBA between the “direct electric field” that solves the induced dipoles versus the “polarization electric field” that is used to compute the electrostatic energy.²⁹ For QM methods like Hartree-Fock (HF) and density functional theory (DFT) where the target state energy gets variationally minimized, their embedding in AMOEBA model faces extra difficulties because the state energy in AMOEBA is no longer variational with respect to the induced dipoles. As a result, a Lagrangian approach must be used to compute their nuclear gradients in AMOEBA,⁴² and this significantly increases the complexity and computational cost compared to embedding such QM method in variational polarizable models like the Drude model.⁴³ Fortunately, because SA-CASSCF variationally minimizes the state-averaged energy, the target state energy is not variational in the first place and the corresponding nuclear gradients already requires the Lagrangian approach regardless of whether AMOEBA is used.⁴⁴ Therefore, the discrepancy between “direct” and “polarization” electric field can be incorporated more easily in the SA-CASSCF framework than HF or DFT.

The second challenge to generalize to protein environment is the treatment of covalent QM/MM boundary and requires interface methods that are accurate for polarizable embedding. Commonly used interface methods include link atom schemes that introduce additional atoms to cap the broken bonds,⁴⁵⁻⁴⁷ and the pseudo-bonding approaches that places an effective core potential at the boundary.^{48,49} Both of these approaches have been successfully applied in the context of embedding in polarizable forcefield including AMOEBA forcefields.^{34,50,51} However, because polarizable embedding of excited state methods in AMOEBA is still far from mature, a careful study of how the approximations made by the QM/MM model affects excited state properties, as well as critical point geometries and conical intersections, is still needed. Such quantitative studies on the errors of different QM/MM models can help decide on a good link atom scheme that is both accurate and conceptually simple.

In this paper, we will present our recent progress that address both of the above two challenges such that our previous work on SA-CASSCF/AMOEBA can now be generalized to covalently bonded macromolecular environments. Section 2 introduces the new theoretical ideas. To address the first challenge about intramolecular polarization, we will start with a brief review of our previous work on SA-CASSCF in AMOEBA water models in Section 2.1 by highlighting which equations depend on the electric field from permanent multipoles. Section 2.2 will then discuss how these working equations should be modified to incorporate direct and polarization electric fields in energy, nuclear gradient, and non-adiabatic coupling calculations respectively. To address the second challenge about covalent QM/MM boundary, Section 2.3 will discuss how the simple single link atom scheme (SL)^{45,46} as well as several variants of the double-link atom scheme (DL)⁴⁷ are applied in SA-CASSCF/AMOEBA. In Section 3, we will present and discuss test results on the new SA-CASSCF/AMOEBA method. Section 3.1 performs detailed quantitative comparisons of the errors incurred by the QM/MM model, and shows that the DL scheme generally yields superior accuracy (relative to full QM) when computing single-point energies, forces, and excitation energies for a benchmark small molecule. In Section 3.2, we also demonstrate the application of the QM/MM model to an elementary reaction in an artificial luciferase, NanoLuc, where we compare the potential energy surfaces for the dissociation of an O-O bonded reaction intermediate of a bioluminescent chromophore in the gas phase, in PCM, and the protein environment described by QM/MM. Our study reveals that the QM/MM environment significantly affects the shape of the PES along the reaction coordinate. Overall, this study is a promising step toward broader studies of photoreactions in biological environments using QM/MM models that couple multireference quantum chemistry methods and polarizable force fields.

2 Method

2.1 Brief review of SA-CASSCF in AMOEBA solvent model without intramolecular polarization

We start with a brief review of polarizable embedding of SA-CASSCF in AMOEBA when intramolecular polarization of MM region is neglected. In the QM region, the electrostatic properties depend on the QM nuclear charges (\mathbf{Z}) and the electronic wavefunction, and the corresponding one-electron density matrix will be denoted as \mathbf{P} . In the MM region, the electrostatic properties are described by the permanent multipoles including charges \mathbf{C} , dipoles \mathbf{M} , and quadrupoles \mathbf{Q} , as well as the induced dipoles \mathbf{d} . Because intramolecular polarization in the MM region mainly affects the electric field generated by the permanent multipoles, we will primarily highlight equations that depend on such electric field $\mathcal{E}^{(\text{perm})}$.

As mentioned in the introduction, our formulation of SA-CASSCF/AMOEBA solves the QM electronic wavefunction and the MM induced dipoles by variationally minimizing the state-averaged energy. Following our previous work, define the ‘‘source’’ as the electronic wavefunction along with all the fixed multipoles (\mathbf{Z} , \mathbf{C} , \mathbf{M} and \mathbf{Q}). The state-averaged energy can then be represented as

$$E^{(\text{SA})} = E^{(\text{SA-source})} - \sum_{Ai} \mathcal{E}_{Ai}^{(\text{SA-source})} d_{Ai} + \frac{1}{2} \sum_{Ai, Bj} d_{Ai} O_{Ai, Bj} d_{Bj} \quad (1)$$

The first term $E^{(\text{SA-source})}$ in Eq.1 is the energy from the source

$$\begin{aligned} E^{(\text{SA-source})} = & \sum_{pq} P_{pq}^{(\text{SA})} K_{pq} + \sum_{pqrs} \Pi_{pqrs}^{(\text{SA})} (pq|rs) \\ & + \sum_{pq} P_{pq}^{(\text{SA})} V_{pq}[\mathbf{Z}, \mathbf{C}, \mathbf{M}, \mathbf{Q}] + \epsilon_{\text{es}}[\mathbf{Z}, \mathbf{C}, \mathbf{M}, \mathbf{Q}] + \epsilon_{\text{vdW}} + \epsilon_{\text{valence}} \end{aligned} \quad (2)$$

where the six terms in Eq.2 represent one electron and two electron interactions, electrostatic interaction between electrons and fixed multipoles, interaction among the fixed multipoles, van der Waals (vdW) interaction and valence interaction respectively. The second term in

Eq.1 describes the interaction of induced dipoles with the electric field generated by the source, which can be further divided into the fields generated by the electrons, QM region nuclear point charges, and the MM region permanent multipoles respectively:

$$\mathcal{E}_{Ai}^{(\text{SA-source})} = \mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{(\text{SA})}] + \mathcal{E}_{Ai}^{(\text{nuc})}[\mathbf{Z}] + \mathcal{E}_{Ai}^{(\text{perm})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}] \quad (3)$$

The last term in Eq.1 describes the mutual interactions among the induced dipoles in the MM region. Given Eq.1, the molecular orbital (MO) coefficient rotations (κ_{rs}), configuration interaction (CI) coefficient rotations (θ_{RS}) and the induced dipoles (d_{Ai}) can then be solved by applying the variational condition to the state-averaged energy:

$$\frac{\partial E^{(\text{SA})}}{\partial \kappa_{rs}} = 0, \frac{\partial E^{(\text{SA})}}{\partial \theta_{RS}} = 0, \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}} = 0 \quad (4)$$

In addition, the dynamic weight scheme further imposes that the weights and energies of states must also satisfy the following self-consistent condition:

$$P_{pq}^{(\text{SA})} = \sum_I w_I P_{pq}^\Theta \quad (5)$$

$$w_I = \omega_{\text{Ref}}(E_I) = \frac{f(|E_I - E_{\text{Ref}}|)}{\sum_K f(|E_K - E_{\text{Ref}}|)} \quad (6)$$

There are various ways to choose the function $f(x)$ (e.g. cubic spline, or sech2 function),⁵² and they all satisfy the property that $f(x)$ approaches 1 if $x \rightarrow 0$, and approaches 0 if $x \geq \Delta$ where Δ is a user selected bandwidth parameter. The energy calculations thus require satisfying four conditions (Eq.4 and Eq.6) simultaneously. Our previous work has implemented a Newton-Raphson based convergence algorithm to solve these conditions, which is adapted from the work of Hohenstein et.al.⁵³ This algorithm requires the first derivatives of $E^{(\text{SA})}$ with respect to each parameter and approximates the Hessian matrix with only the diagonal elements to reduce the computational cost. In particular, the part of first derivative in Eq.4

that contains the source electric field (thus depends on $\vec{\mathcal{E}}^{(\text{perm})}$) is

$$g_{Ai}^{(\text{Q};\text{SA})} = \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}} = \sum_{bj} O_{Ai,Bj} d_{Bj} - \mathcal{E}_{Ai}^{(\text{SA-source})} \quad (7)$$

Finally, once the wavefunction and the induced dipoles are solved from Eq.4, the energy for each state Θ can be computed as

$$E_{\Theta} = E_{\Theta}^{(\text{source})} - \sum_{Ai} \mathcal{E}_{Ai}^{\Theta,(\text{source})} d_{Ai} + \frac{1}{2} \sum_{Ai,Bj} d_{Ai} O_{Ai,Bj} d_{Bj} \quad (8)$$

The state specific source energy $E_{\Theta}^{(\text{source})}$ in Eq.8 is defined by replacing state averaged density $\mathbf{P}^{(\text{SA})}$ in Eq.2 with state specific density \mathbf{P}^{Θ} . Similarly, we can define the state-specific electric field by modifying Eq.3 as

$$\mathcal{E}_{Ai}^{\Theta,(\text{source})} = \mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{\Theta}] + \mathcal{E}_{Ai}^{(\text{nuc})}[\mathbf{Z}] + \mathcal{E}_{Ai}^{(\text{perm})}[\mathbf{C},\mathbf{M},\mathbf{Q}] \quad (9)$$

It is also worth mentioning that because the dynamic weights in Eq.6 only depends on the energy difference between states, the permanent multipole electric field cancels in the difference

$$E_I - E_{\text{ref}} = E_I^{(\text{source})} - E_{\text{Ref}}^{(\text{source})} - \sum_{Ai} \left(\mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^I] - \mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{\text{Ref}}] \right) d_{Ai} \quad (10)$$

This implies that the dynamic weights stay the same whether or not intramolecular polarization is included. To summarize the energy calculations, the permanent multipole electric field $\vec{\mathcal{E}}^{(\text{perm})}$ mainly affects the definitions of the stated averaged energy $E^{(\text{SA})}$ in Eq.1, the state specific energy E_{Θ} in Eq.8, and the first derivative $g_{Ai}^{(\text{Q};\text{SA})}$ in Eq.7.

Because the state specific energy is not variational with respect to the wavefunction coefficients and induced dipoles, the corresponding nuclear gradients need to be determined through a Lagrange multiplier approach. Following our previous work, define the Lagrangian

as

$$L_{\Theta} = E_{\Theta} + \sum_{rs} \bar{\kappa}_{rs}^{\Theta} \frac{\partial E^{(\text{SA})}}{\partial \kappa_{rs}} + \sum_{RS} \bar{\theta}_{RS}^{\Theta} \frac{\partial E^{(\text{SA})}}{\partial \theta_{RS}} + \sum_{Ai} \bar{d}_{Ai}^{\Theta} \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}} + \sum_I \bar{w}_I^{\Theta} \tau_I \quad (11)$$

where $\tau_I = w_I - \omega_{\text{Ref}}(E_I)$ from Eq.6. The Lagrangian introduces four sets of Lagrange multipliers $\bar{\kappa}^{\Theta}$, $\bar{\theta}^{\Theta}$, \bar{d}^{Θ} and \bar{w}^{Θ} , corresponding to constraints on the orbital rotations ('O'), CI rotations ('C'), induced dipoles ('Q'), and dynamic weights ('W') respectively. By imposing the stationary condition on the Lagrangian

$$\frac{\partial L_{\Theta}}{\partial \kappa_{rs}} = 0, \frac{\partial L_{\Theta}}{\partial \theta_{RS}} = 0, \frac{\partial L_{\Theta}}{\partial d_{Ai}} = 0, \frac{\partial L_{\Theta}}{\partial w_I} = 0 \quad (12)$$

one can then solve the Lagrange multipliers from the following coupled-perturbed equation

$$\begin{pmatrix} \mathbb{H}^{(\text{OO})} & \mathbb{H}^{(\text{OC})} & \mathbb{W}^{(\text{OW})} & \mathbb{H}^{(\text{OQ})} \\ \mathbb{H}^{(\text{CO})} & \mathbb{H}^{(\text{CC})} & \mathbf{0} & \mathbb{H}^{(\text{CQ})} \\ \mathbb{H}^{(\text{WO})} & \mathbf{0} & \mathbf{I} & \mathbb{H}^{(\text{WQ})} \\ \mathbb{H}^{(\text{QO})} & \mathbb{H}^{(\text{QC})} & \mathbb{W}^{(\text{QW})} & \mathbb{H}^{(\text{QQ})} \end{pmatrix} \begin{pmatrix} \bar{\kappa}^{\Theta} \\ \bar{\theta}^{\Theta} \\ \bar{d}^{\Theta} \\ \bar{w}^{\Theta} \end{pmatrix} = \begin{pmatrix} -\mathbf{g}^{(\text{O}),\Theta} \\ \mathbf{0} \\ \mathbf{0} \\ -\mathbf{g}^{(\text{Q}),\Theta} \end{pmatrix} \quad (13)$$

The definitions for elements in the above linear systems are provided in our previous work.^{36,37} Because the electric field from permanent multipoles only contributes to the first derivative of the Lagrangian, it does not appear in any of the terms on the left hand side in the coupled perturbed equation. On the right hand side of Eq.13, the only term that contains the source electric field (thus depends on $\vec{\mathcal{E}}^{(\text{perm})}$) is

$$g_{Ai}^{(\text{Q}),\Theta} = \frac{\partial E_{\Theta}}{\partial d_{Ai}} = \sum_{bj} O_{Ai,Bj} d_{Bj} - \mathcal{E}_{Ai}^{(\text{source}),\Theta} \quad (14)$$

Once the Lagrange multipliers are solved, the nuclear gradients can then be computed by taking derivatives of integrals in the Lagrangian, i.e.

$$\frac{dE_{\Theta}}{d\xi} = \frac{\partial L_{\Theta}}{\partial \xi} \quad (15)$$

The detailed expressions are more complicated and can be found in our previous work.^{36,37} In particular, there are two terms in the Lagrangian L_Θ that depend on $\vec{\mathcal{E}}^{(\text{perm})}$. One is from the first term E_Θ that contains $-\sum_{Ai} d_{Ai} \mathcal{E}_{Ai}^{(\text{perm})}$ based on Eq.8. The other is from the fourth term $\sum_{Ai} \bar{d}_{Ai}^\Theta \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}}$ in L_Θ that contains $-\sum_{Ai} \bar{d}_{Ai}^\Theta \mathcal{E}_{Ai}^{(\text{perm})}$ based on Eq.7. By combining these two, the contribution from the permanent multipole electric field to the nuclear gradient is

$$W_\Theta^\xi = - \sum_{Ai} (d_{Ai} + \bar{d}_{Ai}^\Theta) \frac{\partial \mathcal{E}_{Ai}^{(\text{perm})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]}{\partial \xi} \quad (16)$$

To summarize the gradient calculations, the permanent multipole electric field $\vec{\mathcal{E}}^{(\text{perm})}$ mainly occurs at two places: one is the $g_{Ai}^{(\text{Q}),\Theta}$ in Eq.14 that appears on the right hand side of the coupled-perturbed equation, another is W_Θ^ξ in Eq.16 that contributes to the final nuclear gradients.

The non-adiabatic coupling between two states Θ and Π can be computed as

$$\chi_\xi^{\Theta\Pi} = \delta_\xi^{\Theta\Pi} - \frac{1}{E_\Theta - E_\Pi} \frac{d\Omega_{\Theta\Pi}}{d\xi} \quad (17)$$

where $\delta_\xi^{\Theta\Pi}$ can be computed from derivative of overlap integrals, and $\Omega_{\Theta\Pi}$ can be computed from the transition density matrices as

$$\Omega_{\Theta\Pi} = \sum_{pq} T_{pq}^{\Theta\Pi} \eta_{pq} + \sum_{pqrs} \Pi_{pqrs}^{\Theta\Pi} (pq|rs) \quad (18)$$

Similar to the analytical nuclear gradients, $\frac{d\Omega_{\Theta\Pi}}{d\xi}$ can be computed through the Lagrange multiplier approach by defining the Lagrangian as

$$L_{\Theta\Pi} = \Omega_{\Theta\Pi} + \sum_{rs} \bar{\kappa}_{rs}^{\Theta\Pi} \frac{\partial E^{(\text{SA})}}{\partial \kappa_{rs}} + \sum_{RS} \bar{\theta}_{RS}^{\Theta\Pi} \frac{\partial E^{(\text{SA})}}{\partial \theta_{RS}} + \sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}} + \sum_I \bar{w}_I^{\Theta\Pi} \tau_I \quad (19)$$

$L_{\Theta\Pi}$ should also satisfy the stationary condition Eq.12. Because $\Omega_{\Theta\Pi}$ in Eq.18 does not contain the permanent multipole electric field, none of the terms in the coupled-perturbed

equation for NAC calculations depends on $\vec{\mathcal{E}}^{(\text{perm})}$. As a result, when computing nuclear derivatives through $\frac{d\Omega_{\Theta\Pi}}{d\xi} = \frac{\partial L_{\Theta\Pi}}{\partial \xi}$, the only term that depends on the permanent multipole electric field is $\sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \frac{\partial E^{(\text{SA})}}{\partial d_{Ai}}$ that contains $-\sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \mathcal{E}_{Ai}^{(\text{perm})}$. Therefore, the permanent multipole electric field will contribute to $\frac{d\Omega_{\Theta\Pi}}{d\xi}$ in the NAC through the following quantity:

$$W_{\Theta\Pi}^{\xi} = - \sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \frac{\partial \mathcal{E}_{Ai}^{(\text{perm})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]}{\partial \xi} \quad (20)$$

In summary, NAC calculations only depend on the permanent multipole electric field $\vec{\mathcal{E}}^{(\text{perm})}$ through one single term, i.e. $W_{\Theta\Pi}^{\xi}$ in Eq.20.

2.2 Including intramolecular polarization

When intramolecular polarization is included in the AMOEBA model, the main difference is that there are two types of electric field generated from the permanent multipoles of the MM atoms, and they mainly differ in how the short range interactions are scaled. The differences in the scaling factors and motivations for such design are explained in the original AMOEBA papers.^{27,29} One is the ‘‘direct field’’ $\vec{\mathcal{E}}^{(\text{perm-D})}$ that is used to determine the values of the induced dipoles. The scaling factors of direct field involve partitioning the molecule into fragments called ‘‘polarization groups’’, and are designed to prevent the formation of unphysically large induced dipoles. The second type is the ‘‘polarization field’’ $\vec{\mathcal{E}}^{(\text{perm-P})}$ that is used to compute the energy. The scaling factors for polarization field follow similar rules as other force fields including fixed charge models, and primarily depend on the number of covalent bonds that separate a pair of atoms.

First consider the energy calculations. Because the values of the induced dipoles are determined from the state averaged energy $E^{(\text{SA})}$ (see Eq.4), the permanent electric field in $E^{(\text{SA})}$ should be modified using the ‘‘direct field’’, i.e.

$$\mathcal{E}_{Ai}^{(\text{SA-source-D})} = \mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{(\text{SA})}] + \mathcal{E}_{Ai}^{(\text{nuc})}[\mathbf{Z}] + \mathcal{E}_{Ai}^{(\text{perm-D})}[\mathbf{C}, \mathbf{D}, \mathbf{Q}] \quad (21)$$

$$E^{(\text{SA-D})} = E^{(\text{SA-source})} - \sum_{Ai} \mathcal{E}_{Ai}^{(\text{SA-source-D})} d_{Ai} + \frac{1}{2} \sum_{Ai,Bj} d_{Ai} O_{Ai,Bj} d_{Bj} \quad (22)$$

When solving the wavefunction coefficients and the induced dipoles using Newton-Raphson algorithm, the first derivatives of Eq.7 should thus be modified correspondingly using the direct field as

$$g_{Ai}^{(\text{Q;SA-D})} = \frac{\partial E^{(\text{SA-D})}}{\partial d_{Ai}} = \sum_{bj} O_{Ai,Bj} d_{Bj} - \mathcal{E}_{Ai}^{(\text{SA-source-D})} \quad (23)$$

The state averaged energy $E^{(\text{SA-D})}$ in Eq.22 comes from averaging the state specific energy of all states using the direct field, i.e. $E_{\Theta}^{(\text{D})}$. However, instead of directly reporting $E_{\Theta}^{(\text{D})}$ as the final results, we now need an additional step to incorporate the effect from the ‘‘polarization field’’ into the state specific energy $E_{\Theta}^{(\text{P})}$. In pure MM AMOEBA calculation, the total energy using the direct field is different from using the polarization field by

$$\Delta E^{(\text{P-D})} = -\frac{1}{2} \sum_{Ai} \left(\mathcal{E}_{Ai}^{\Theta,(\text{Perm-P})} - \mathcal{E}_{Ai}^{\Theta,(\text{Perm-D})} \right) d_{Ai} \quad (24)$$

Therefore, we require that state energies in SA-CASSCF/AMOEBA should continue satisfying $E_{\Theta}^{(\text{P})} = E_{\Theta}^{(\text{D})} + \Delta E^{(\text{P-D})}$, and this leads to the following equations

$$\mathcal{E}_{Ai}^{\Theta,(\text{source-PD})} = \mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{\Theta}] + \mathcal{E}_{Ai}^{(\text{nuc})}[\mathbf{Z}] + \frac{1}{2} \left(\mathcal{E}_{Ai}^{(\text{perm-P})} + \mathcal{E}_{Ai}^{(\text{perm-D})} \right) [\mathbf{C}, \mathbf{D}, \mathbf{Q}] \quad (25)$$

$$E_{\Theta}^{(\text{P})} = E_{\Theta}^{(\text{source})} - \sum_{Ai} \mathcal{E}_{Ai}^{\Theta,(\text{source-PD})} d_{Ai} + \frac{1}{2} \sum_{Ai,Bj} d_{Ai} O_{Ai,Bj} d_{Bj} \quad (26)$$

Note that $\mathcal{E}_{Ai}^{\Theta,(\text{source-PD})}$ depends on both the polarization field $\mathcal{E}_{Ai}^{(\text{perm-P})}$ and the direct field $\mathcal{E}_{Ai}^{(\text{perm-D})}$.

For gradient calculations, we first need to consider how to modify the Lagrangian. In Eq.11, the first term is the target state energy thus should now be replaced with Eq.26. The remaining terms constrain how the wavefunctions and induced dipoles are determined, thus

should be replaced with Eq.22. Therefore, the Lagrangian should be modified as follows

$$\tilde{L}_\Theta = E_\Theta^{(P)} + \sum_{rs} \bar{\kappa}_{rs}^\Theta \frac{\partial E^{(SA-D)}}{\partial \kappa_{rs}} + \sum_{RS} \bar{\theta}_{RS}^\Theta \frac{\partial E^{(SA-D)}}{\partial \theta_{RS}} + \sum_{Ai} \bar{d}_{Ai}^\Theta \frac{\partial E^{(SA-D)}}{\partial d_{Ai}} + \sum_I \bar{w}_I^\Theta \tau_I \quad (27)$$

The Lagrangian should continue to satisfy the stationary condition in Eq.12, which leads to the modified coupled-perturbed equation. Recall that Eq.14 is the only term in the coupled-perturbed equation that depends on the permanent multipole electric field. Since it corresponds to first derivatives of the target state energy, it should be modified using $E_\Theta^{(P)}$ as

$$g_{Ai}^{(Q-P),\Theta} = \frac{\partial E_\Theta^{(P)}}{\partial d_{Ai}} = \sum_{bj} O_{Ai,Bj} d_{Bj} - \mathcal{E}_{Ai}^{(\text{source-PD}),\Theta} \quad (28)$$

Finally, the nuclear gradients Eq.15 should be computed now using the modified Lagrangian Eq.27, and it depends on the permanent multipole electric field at two places: the first term $E_\Theta^{(P)}$ (Eq.26) that contains $-\frac{1}{2} \sum_{Ai} d_{Ai} \left(\mathcal{E}_{Ai}^{(\text{perm-P})} + \mathcal{E}_{Ai}^{(\text{perm-D})} \right)$, and the fourth term $\sum_{Ai} \bar{d}_{Ai}^\Theta \frac{\partial E^{(SA-D)}}{\partial d_{Ai}}$ that contains $-\sum_{Ai} \bar{d}_{Ai}^\Theta \mathcal{E}_{Ai}^{(\text{perm-D})}$ (Eq.23). Therefore, the nuclear gradient contributions from the permanent multipole electric field in Eq.16 should be modified correspondingly as

$$\tilde{W}_\Theta^\xi = -\frac{1}{2} \sum_{Ai} d_{Ai} \frac{\partial \mathcal{E}_{Ai}^{(\text{perm-P})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]}{\partial \xi} - \sum_{Ai} \left(\frac{1}{2} d_{Ai} + \bar{d}_{Ai} \right) \frac{\partial \mathcal{E}_{Ai}^{(\text{perm-D})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]}{\partial \xi} \quad (29)$$

Note that in pure classical AMOEBA calculations, even though the energy is not variational with respect to the induced dipoles, the nuclear gradients can still be computed through the chain rule²⁹ and the Lagrange multiplier approach is not necessary. However, using chain rule to compute QM/MM nuclear gradients in the non-variational situation is generally infeasible, because it is difficult to obtain partial derivatives of wavefunction coefficients with respect to nuclear positions due to the complicated dependencies. The Lagrange multiplier approach essentially avoids directly evaluating these partial derivatives by reformulating the dependencies as constraints that the wavefunction coefficients need to satisfy.

Similarly, for non-adiabatic couplings, the Lagrangian in Eq.19 for computing $\frac{d}{d\xi}\Omega_{\Theta\Pi}$ should be modified as

$$\tilde{L}_{\Theta\Pi} = \Omega_{\Theta\Pi} + \sum_{rs} \bar{\kappa}_{rs}^{\Theta\Pi} \frac{\partial E^{(\text{SA-D})}}{\partial \kappa_{rs}} + \sum_{RS} \bar{\theta}_{RS}^{\Theta\Pi} \frac{\partial E^{(\text{SA-D})}}{\partial \theta_{RS}} + \sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \frac{\partial E^{(\text{SA-D})}}{\partial d_{Ai}} + \sum_I \bar{w}_I^{\Theta\Pi} \tau_I \quad (30)$$

No modification is needed for the coupled perturbed equation since it does not contain the permanent multipole electric field. Once the Lagrange multipliers are solved, contribution to $\frac{d}{d\xi}\Omega_{\Theta\Pi}$ from the permanent multipole electric field in Eq.20 should be modified correspondingly as

$$\tilde{W}_{\Theta\Pi}^{\xi} = - \sum_{Ai} \bar{d}_{Ai}^{\Theta\Pi} \frac{\partial \mathcal{E}_{Ai}^{(\text{perm-D})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]}{\partial \xi} \quad (31)$$

In summary, to include intramolecular polarization effects, one can reuse the algorithm described in our previous work^{36,37} with a few modifications. For energy calculations, (1) during the Newton-Raphson algorithm, the gradient Eq.7 should be replaced with Eq.23, and (2) once the algorithm is converged, the state energy should be recomputed using Eq.26. For gradient calculations, (1) the gradient Eq.14 passed to the coupled perturbed solver should be replaced with Eq.28, and (2) the contribution to nuclear gradients from the permanent multipole electric field Eq.16 should be replaced with Eq.29. For non-adiabatic coupling calculations, the coupled perturbed solver does not need any modifications, and the contribution to nuclear gradients from the permanent multipole electric field Eq.20 should be replaced with Eq.31.

2.3 QM/MM Link atom

In this section, we will discuss how to address the QM/MM boundary when it cuts through a covalent bond, which is the second challenge for applying SA-CASSCF/AMOEBA to protein environment. When the QM/MM boundary cuts through a covalent bond, the distances between the MM charges and the QM region can become very short, potentially giving rise to unphysically strong electrostatic and polarization interactions. Moreover, the

MM region is often left with a non-integer total charge, as well as bonded interactions that include both QM and MM atoms, which need to be treated carefully in the setting up of the QM/MM model. Given the SA-CASSCF/AMOEBA energy expression in Eq.22 and Eq.26, the different energy components can be divided into the electrostatic energy as well as the non-electrostatic energy. The treatments of electrostatic interaction and non-electrostatic interaction will be discussed in Section 2.3.1 and 2.3.2 respectively.

2.3.1 Electrostatic interaction energy

Because the state-averaged energy $E^{(\text{SA-D})}$ in Eq.22 and the state specific energy $E_{\Theta}^{(\text{P})}$ in Eq.26 are closely related, to keep the expressions simple, we will explain the concepts using $E^{(\text{SA-D})}$, and similar concepts can be easily generalized to $E_{\Theta}^{(\text{P})}$. The electrostatic interaction energy in SA-CASSCF/AMOEBA can be further divided into the following components: interaction within the QM region, interaction within the MM region, and interaction between the QM and MM regions. One general principle we follow in this work is that different quantities on the same atom (e.g. electrons and nuclear charges on the same QM atom, permanent multipoles and induced dipoles on the same MM atoms) will always experience the same electrostatic interaction.

The interaction within the QM region in Eq.22 includes the interaction among the the electronic wavefunction and the nuclear point charges \mathbf{Z}

$$E^{(\text{SA-QM})} = \sum_{pq} P_{pq}^{(\text{SA})} K_{pq} + \sum_{pqrs} \Pi_{pqrs}^{(\text{SA})}(pq|rs) + \sum_{pq} P_{pq}^{(\text{SA})} V_{pq}[\mathbf{Z}] + \epsilon_{\text{es}}[\mathbf{Z}] \quad (32)$$

Note that although the electron kinetic energy does not rigorously belong to electrostatic interaction, we still include it in $E^{(\text{SA-QM})}$ since it only exists in the QM region. In pure quantum mechanical SA-CASSCF calculations, electrons and nuclear charges always interact with each other through the full Coulomb interaction $\frac{1}{r}$. As a result, we choose that Coulomb interaction within the QM region is always described with $\frac{1}{r}$.

The interaction within the MM region among the permanent multipoles and induced dipoles include the following terms

$$E^{(\text{MM-D})} = \epsilon_{\text{es}}[\mathbf{C}, \mathbf{M}, \mathbf{Q}] - \sum_{Ai} \mathcal{E}_{Ai}^{(\text{perm-D})}[\mathbf{C}, \mathbf{M}, \mathbf{Q}]d_{Ai} + \frac{1}{2} \sum_{Ai, Bj} d_{Ai} O_{Ai, Bj} d_{Bj} \quad (33)$$

Because this is pure MM interaction, we choose to treat them in exactly the same way as in pure MM AMOEBA force field. In pure MM models, it is common practice to define exclusion rules for each MM multipole, such that each one experiences its “own” electric field that comes from only those “other” multipoles that are separated by a sufficient number of bonds. These exclusion rules are specified by the covalent maps (denoted as $\text{CovMap}(\frac{1}{r})$). There are two types of covalent maps in the AMOEBA models : the “bond covalent maps” describes the number of covalent bonds that separate every pair of atoms, and the “polarization covalent maps” describes the number of polarization groups that separate every pair of atoms. When computing the electric field from the permanent multipoles, the reason why we have direct and polarization electric fields is because they are computed using covalent maps that differ in their scaling factors. Since the covalent maps are initially constructed with the assumption that all atoms are treated with AMOEBA models, the atoms that actually belong to the QM region should be deleted from covalent maps of any MM atoms.

Finally, the most complicated part is the interaction between the QM region (electron density \mathbf{P} and nuclear charge \mathbf{Z}) and MM region (permanent multipoles $\mathbf{C}, \mathbf{M}, \mathbf{Q}$ and the induced dipoles \mathbf{d}):

$$E^{(\text{SA-QM-MM})} = \sum_{pq} P_{pq}^{\ominus} V_{pq}[\mathbf{C}, \mathbf{M}, \mathbf{Q}] + \epsilon_{\text{es}}^{(\text{QM-MM})}[\mathbf{Z}, \{\mathbf{C}, \mathbf{M}, \mathbf{Q}\}] - \sum_{Ai} \left(\mathcal{E}_{Ai}^{(\text{elec})}[\mathbf{P}^{\ominus}] + \mathcal{E}_{Ai}^{(\text{nuc})}[\mathbf{Z}] \right) d_{Ai} \quad (34)$$

The treatment of $E^{(\text{SA-QM-MM})}$ will depend on the link atom scheme. In this work, we have considered the single link atom scheme (SL) and the double link atom scheme (DL) as discussed below.

Single link atom scheme (SL)

Table 1: Summary of how the single link (SL) atom scheme is set up. For atoms in each region, the middle columns explain what types of electrostatic quantities the corresponding atoms possess. The right column summarizes starting from the purely classical AMOEBA system, what updates are needed for each region in order to set up the corresponding QM/MM system.

Region in SL	Electrostatic Quantities	Setup
QM	Point nuclear charges \mathbf{Z} Wavefunction	Update point charges to nuclear charges
QMLink	Point nuclear charges \mathbf{Z} Wavefunction	Set nuclear charge and basis set to Hydrogen atom
MMHost	None	Set permanent multipoles and induced dipoles to 0 Delete QM atoms from covalent maps
Other MM	Permanent multipoles $\mathbf{C}, \mathbf{M}, \mathbf{Q}$ Induced dipoles \mathbf{d}	Evenly spread Δq from above updates to all MM non host atoms Delete QM atoms from covalent maps

Table 2: Summary of the choices of electrostatic interactions between different regions in the single link (SL) atom scheme. Because the table is symmetric, only the elements in the upper triangle are provided. The asterisk "*" indicates that different options are provided for the corresponding entry, and can be selected by the user. In the software column, "QC" indicates that the corresponding terms are implemented in the quantum chemistry software (Terachem in our implementation), and the "MM" indicates that the corresponding terms are computed by the molecular dynamics software (OpenMM in our implementation).

Region in SL	QM	QMLink	MMHost	Other MM	Software
QM	$\frac{1}{r}$	$\frac{1}{r}$	0	* $\frac{1}{r}$ or $\frac{\text{erf}(\omega_{MM}r)}{r}$	QC
QMLink		$\frac{1}{r}$	0	* $\frac{1}{r}$ or $\frac{\text{erf}(\omega_{MM}r)}{r}$	
MMHost			0	0	
Other MM				CovMap($\frac{1}{r}$)	MM

When a covalent bond A-B is cut by the QM/MM boundary such that atom A is in the QM region and atom B is in the MM region, then we will call atom A the "QMHost" and atom B the "MMHost". In the single link atom approach, a hydrogen atom will be attached to QMHost atom A as "A-Hq", where the "Hq" atom will be called the "QMLink" atom. The bond length of A-Hq will be fixed to 1.09 Angstrom, and is parallel to the direction of A-B bond. The QMLink atom is treated as part of the QM region, and adopts nuclear charge and basis functions of a hydrogen atom.

In addition to the insertion of QMLink atom, the properties of MM atoms still need to be updated to properly set up the QM/MM system. On one hand, the short distance

between QMLink atom and MMHost atom will likely lead to excessively large electrostatic interaction. The common strategy when applying SL scheme to fixed charge force field is to simply zero out the charge on the MMHost. Following the same idea, we also choose to remove all permanent multipoles and induced dipoles from the MMHost atoms. On the other hand, a small charge discrepancy Δq often gets accumulated due to the change in charges of the QM region, QMLink and MMHost atoms. We choose to evenly distribute Δq to all the remaining MM atoms, such that the overall system has the correct charge.

Finally, we need to choose how the QM region (including QMLink) interacts with the MM region (excluding MMHost) in Eq.34. In this work, we have tested two different choices. One choice is to let QM region interact with the MM region through the full Coulomb interaction $\frac{1}{r}$. This is what we have implemented in our previous work. A second choice is to let them interact through a damped Coulomb interaction $\frac{\text{erf}(\omega_{\text{MM}}r)}{r}$, with ω_{MM} being a user chosen parameter. Physically, using a damped Coulomb interaction is equivalent to delocalizing the point multipoles on MM atoms when they interact with the QM region. The smaller the value of ω_{MM} is, the more delocalized the multipoles become, and the more the Coulomb interaction gets damped in the short range. The molecular integrals corresponding to the damped Coulomb interaction are quite simple to obtain using the existing codes based on the McMurchie-Davidson algorithm.⁵⁴ One only needs to modify the fundamental integral, while the recursive algorithm for increment of angular momentum stays the same.

The setup of SA-CASSCF/AMOEBa using single-link atom scheme is summarized in Table 1, and our choices of the electrostatic interactions between different regions are summarized in Table 2. One thing to note is that we have changed how different terms are distributed between QM software and MM software. In our previous work, we let QM software handle terms that involve electrons, while the remaining terms (including QM nuclear charges) are all handled by the MM software. However, we realize that our previous strategy is not capable of applying custom MM electrostatic interactions consistently to the QM electrons and QM nuclear charges, since they are computed by different codes. For exam-

ple, there is no simple way to use the standard OpenMM APIs to compute $\frac{\text{erf}(\omega r)}{r}$ type of electrostatic interactions between the QM nuclear charges and MM region. To prevent this potential inconsistency, we change to let QM software handle terms that involve QM atoms including electrons and QM nuclear charges in this work, and this new design principle is also reflected in Table 2. It is also worth pointing out that there are other ways to set up the single link atom scheme for AMOEBA polarizable embedding as reported in other works. For example, Nottoli et.al³⁴ choose to zero out QM–MM electrostatic interactions for all the MM atoms within 2 or 3 bonds from the QM region, and the charge discrepancy Δq is only distributed to the neighboring heavy atoms. Therefore, the performance of the SL scheme may vary between our implementation and others.

Double link atom scheme (DL)

Table 3: Summary of how the double link atom scheme is set up. The table is formatted in the same way as Table 1. The asterisk ”*” indicates that different options are provided for the corresponding entry, and can be selected by the user.

Region in DL	Electrostatic Quantities	Setup
QM	Point nuclear charges Z Wavefunction	Update point charges to nuclear charges
QMLink	Point nuclear charges Z Wavefunction	Set nuclear charge and basis set to Hydrogen atom
MMLink	Permanent charges C	Evenly spread Δq from above to all MMLink atoms Set bond covalent map to be the same as the QMHost Set polarization covalent map to be the same as MMHost
MMHost	Permanent charges C *Permanent M,Q or None *Induced dipoles d or None	* Redistribute permanent charges or stay the same In bond covalent maps, replace QMHost with MMLink
Other MM	Permanent multipoles C,M,Q Induced dipoles d	In polarization covalent maps, insert MMLink if MMHost is present Delete remaining QM atoms from all maps

Table 4: Summary of the choices of electrostatic interactions between different regions in the double link (DL) atom scheme. The table is formatted in the same way as Table 2.

Region in DL	QM	QMLink	MMLink	MMHost	Other MM	software
QM	$\frac{1}{r}$	$\frac{1}{r}$	$\frac{\text{erf}(\omega_{\text{link}} r)}{r}$	$\frac{\text{erf}(\omega_{\text{link}} r)}{r}$	* $\frac{1}{r}$ or $\frac{\text{erf}(\omega_{\text{MM}} r)}{r}$	QC
QMLink		$\frac{1}{r}$	$\frac{\text{erf}(\omega_{\text{link}} r)}{r}$	$\frac{\text{erf}(\omega_{\text{link}} r)}{r}$	* $\frac{1}{r}$ or $\frac{\text{erf}(\omega_{\text{MM}} r)}{r}$	
MMLink			CovMap($\frac{1}{r}$)	CovMap($\frac{1}{r}$)	CovMap($\frac{1}{r}$)	MM
MMHost				CovMap($\frac{1}{r}$)	CovMap($\frac{1}{r}$)	
Other MM					CovMap($\frac{1}{r}$)	

In the double link atom scheme, a QMLink atom is inserted and set up in the same

way as the SL scheme. One drawback of the SL scheme is that although the QM region can be considered as a complete molecule, the MM region still contains broken bonds at the QM/MM boundary. Therefore, the DL scheme introduces a second hydrogen atom that is attached to the MMHost atom B as “Hm-B”, where the “Hm” atom will be called the “MMLink” atom. Similar to the QMLink, the bond length of Hm-B is also fixed at 1.09 angstrom along the direction of the QMHost-MMHost bond. The MMLink is considered as part of the MM region, and it caps the broken bond such that the MM region also behaves like a complete molecule.

Because the MMLink atom is added into the MM region, the corresponding updates needed for MM atoms are quite different from SL scheme. First, instead of distributing the charge discrepancy Δq accumulated from the QM region to all MM atoms, we choose to only distribute Δq to the MMLink atoms. This prevents modifying the electrostatic interactions of the MM region globally, which may better preserve the quality of the force field. Second, unlike the SL scheme, permanent multipoles and induced dipoles on the MMHost don’t have to be zeroed out in the DL scheme. Regarding the permanent charges \mathbf{C} on the MMHost, we have implemented the option of either keeping the charge as is, or applying the “redistribution algorithm” and its details are described in the SI-Section 1. The redistribution algorithm attempts to transfer charge from the MMHost to directly bonded MM atoms with opposite charges, which achieves an overall reduction in the magnitude of charge without modifying charges elsewhere in the system. Furthermore, we also provide the options allowing the user to choose whether permanent dipoles/quadrupoles on the MMHost are kept or not, and whether induced dipoles on the MMHost are kept or not. Effects from different combinations of these options will be studied in Section 3.1.

In addition, because the MMLink atom is added to the MM region through a covalent bond, the covalent maps must be updated properly in order to correctly calculate the electrostatic interaction in the pure MM region (Eq.33). On one hand, for the “bond covalent maps” that describe the bonding patterns, the MMLink atom takes up the position of the

QMHost atom and is covalently bonded to the MMHost atom. Therefore, MMLink atom should inherit the bond covalent maps (keeping only MM atoms) of the QMHost atom. At the same time, if any MM atom used to contain QMHost in its bond covalent maps indicating that it is separated from QMHost by n bonds, the QMHost should now be replaced with the MMLink atom. On the other hand, for the “polarization covalent maps” that describe the connectivity of polarization groups, the MMLink atom should always belong to the same polarization group as the MMHost atom. As a result, MMLink atom should inherit the polarization covalent maps from the MMHost atom. At the same time, if any other MM atom contains MMHost in its polarization group maps, the MMLink atom needs to be inserted into the same map.

Finally, we need to consider the electrostatic interaction between the QM and MM region. Here we separate the MM region into two groups: one group contains the MMLink and MMHost atoms, and a second group contains the remaining MM atoms. For the first group, because the MMLink and MMHost atoms are located so close to the QM region, to prevent overpolarization, we let them always interact through damped Coulomb interactions $\frac{\text{erf}(\omega_{\text{link}}r)}{r}$ where ω_{link} can be chosen by the user. For the second group that contains the remaining MM atoms, their interaction with the QM region is treated in the same way as the SL scheme, i.e. either the full Coulomb interaction $\frac{1}{r}$ or damped Coulomb interaction $\frac{\text{erf}(\omega_{\text{MM}}r)}{r}$.

The setup of SA-CASSCF/AMOEBA using double-link atom scheme is summarized in Table 3, and the electrostatic interactions between different regions are summarized in Table 4. Similar to the SL scheme, terms involving QM electrons and QM nuclei are handled by the QM software, while only pure MM interactions are handled by the MM software.

2.3.2 Non-electrostatic energy $\epsilon_{\text{valence}}$ and ϵ_{vdW}

The link atom schemes described in Section 2.3.1 are introduced for the purpose of describing electrostatic interactions across the QM/MM boundary. As a result, the QMLink atoms and MMLink atoms do not explicitly participate in other interactions such as the valence term

and the vdW interactions.

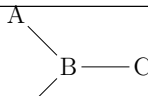
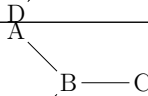
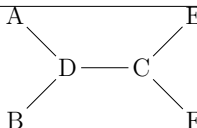
In general, the QM method is responsible for capturing the vdW interactions among the QM atoms, and the AMOEBA model is responsible for capturing the vdW interactions between the QM and MM regions, as well as the interactions among the MM atoms. Therefore, when setting up the QM/MM system, any MM vdW interaction between a pair of QM atoms is deleted. The MM vdW term approximates both the exchange repulsion and dispersion interactions. When the QM region is described with SA-CASSCF, it is able to capture exchange repulsion but is relatively poor at describing dispersion due to the lack of dynamic correlation. The dispersion in the QM region will be addressed by applying perturbation theory on top of SA-CASSCF in future works, thus we do not add any dispersion correction for the QM region in this work.

The valence energy terms, which include bond stretching, angle bending, torsion, out-of-plane bending, pi-torsion, and torsion-torsion, generally involve 2 to 6 atoms that are connected in various ways in the MM topology. In the simplest situation, a given energy term that only involves QM atoms is deleted, and those that only involve MM atoms are kept. However, for energy terms near the QM/MM boundary that involve both QM and MM atoms, it is less obvious whether the term should be kept or not. In this work, the decision of whether to delete an MM energy term is based on whether the corresponding interaction has already been adequately captured by the link atom. For example, we keep the bond term between QMHost and MMHost to model the corresponding bond stretching, because when inserting the link atom, the bond length of QMHost-QMLink (as well as MMHost-MMLink) is defined to be constant thus is unaffected by the bond length of QMHost-MMHost. On the other hand, we delete the angle energy term for the angle bending of QM-QMHost-MMHost. This is because QMHost-QMLink is parallel to the QMHost-MMHost bond, and therefore the angle bending effect will be approximated by the bending of QM-QMHost-QMLink within the QM region. Valence terms involving more particles such as torsion terms are intended to model a mixture of covalent and noncovalent interactions. The electrostatic

portion of these noncovalent interactions are already included in the QM/MM Hamiltonian, therefore the choices to keep or remove torsion and higher terms are largely motivated by physical intuition. For example, we choose to keep the MM torsion term for (QM-2)-(QM-1)-QMHost-MMHost, because although (QM-2)-(QM-1)-QMHost-QMLink shares the same torsion angle, the steric contributions that the MM term accounts for will not be included in the QM energy as QMLink is a hydrogen atom and not sterically hindered. Overall, we follow the conventions of Das et al.⁴⁷ for valence terms that AMOEBA shares with the AMBER & CHARMM models, and similar conventions are proposed here for the pi-torsion and torsion-torsion terms that are specific to AMOEBA.

Based on the discussions in this section, for every type of non-electrostatic interaction in AMOEBA force field, we have summarized the corresponding condition that determines whether a energy term should be deleted or not in Table 5.

Table 5: Summary of how non-electrostatic interactions are treated when the atoms involved cross the QM/MM boundary. The names in the "Force" column follow the class names in OpenMM. The "Bonding" column illustrates how the atoms involved in a particular force are bonded to each other. The last column provides the condition to remove a force component based on the bonding pattern.

Number of Atoms	Force	Bonding	Condition to Remove from Classical Force
2	AmoebaBond	A-B	Both A and B are QM
	HarmonicBond	A-B	Both A and B are QM
3	AmoebaAngle	A-B-C	B is QM
	StretchBend	A-B-C	B is QM
4	PeriodicTorsion	A-B-C-D	Both B and C are QM
	InPlaneAngle		B is QM
	OutOfPlaneBend		B is QM
5	TorsionTorsion	A-B-C-D-E	B and C and D are all QM
6	PiTorsion		Both C and D are QM
Long-Range	Vdw	A...B	Both A and B are QM

3 Results and discussion

The method described in this work is implemented in the TeraChem^{55,56} quantum chemistry package and is interfaced to a modified version of the OpenMM software version 8.1⁵⁷ (using the “Reference” platform) for computing terms related to the AMOEBA force field.⁵⁸ All calculations are performed on computing nodes with NVIDIA GeForce GTX 1080Ti GPUs and Intel Xeon E5-2637 CPUs. Figures and tables in the supporting information (SI) will be referred to as SI-Figures and SI-Tables.

3.1 Comparison of various options in link atom schemes

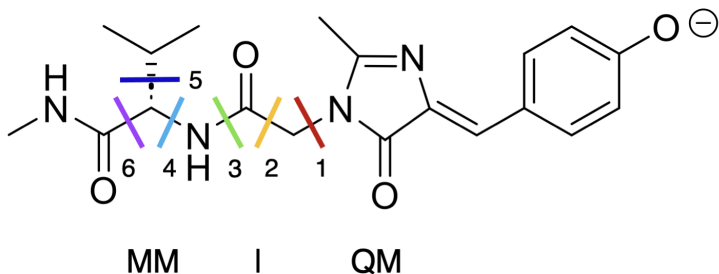


Figure 1: Chemical structure of the molecule used for testing different link atom scheme. The molecule consists of anionic *p*-hydroxybenzylidene-imidazolidone chromophore (HBDI, found in green fluorescent protein) covalently linked to valine. Solid lines in different colors along with the numbers indicate six different choices of QM region used in the tests (the QM region always includes the chromophore on the right).

To compare different link atom schemes, we design the following test system. The molecule we used is illustrated in Figure 1, which is *p*-hydroxybenzylidene-imidazolidone (“HBDI”, chromophore in green fluorescent protein) covalently linked to valine. In order to set up QM/MM calculations, we first parametrize the AMOEBA force field for the entire molecule using SA(2)-CASSCF(4e, 3o)/3-21G using the procedure described in SI-Section 4, and the software involved include Gaussian,⁵⁹ Tinker 8,⁶⁰ MultiWfn,⁶¹ and ForceBalance.^{62–64} The parameters are also provided in the SI.

In order to create samples of molecular structures, we energy-minimized two conformations (corresponding to two local energy minima of the valine residue) and each one was used to initiate an ab initio molecular dynamics simulation in the NVE ensemble for 20 ps. Initial velocities were taken from the Maxwell-Boltzmann distribution at 1000 K and a time step of 0.5 fs was used. We then sample the structures every 80 frames (40 ps), leading to 1000 sampled geometries that will be used for testing.

As illustrated in Figure 2, we tested on six different choices of QM region for each conformation. The chosen QM/MM boundaries cover various possibilities when applying QM/MM to protein, including C-C single bond of side chain (QM5, deep blue), amide bonds of backbone (QM3, green), C $_{\alpha}$ -N bonds along the backbone (QM1 red and QM4, cyan), and C $_{\alpha}$ -C bonds along the backbone (QM2 yellow and QM6 purple). The QM region is anionic in all cases.

Testing of parameters on ground state properties

We start with testing the accuracy of the link atom schemes by comparing energies and gradients from QM/MM calculations to full QM calculations. The full QM calculations are computed by SA(2)-CASSCF(4e,3o)/3-21G, and the QM/MM calculations are computed using the same level of quantum theory embedded in AMOEBA model. As described in the previous paragraph, the model in these tests uses the AMOEBA functional form but the parameters are fitted to SA(2)-CASSCF(4e, 3o)/3-21G quantum calculations. Because the quantum theory used in the calculations is consistent with how the AMOEBA models are parametrized, the difference between QM/MM and full QM calculations will be primarily related to the choice of link-atom scheme.

Figure 2 presents the distribution of errors using various settings of the link atom schemes. The error in the energy for one particular structure is calculated as the absolute error in the relative energy between QM/MM and full QM calculations, i.e. $\left| E_{\text{relative}}^{(\text{QM})} - E_{\text{relative}}^{(\text{QM/MM})} \right|$, where the relative energies are computed as $E_{\text{relative}}^{(\text{QM})} = E^{(\text{QM})} - \langle E^{(\text{QM})} \rangle$ and $E_{\text{relative}}^{(\text{QM/MM})} = E^{(\text{QM/MM})} - \langle E^{(\text{QM/MM})} \rangle$. The error in the gradient for one structure is calculated by taking

the maximum value of the norms of the gradient differences on each atom in the molecule $|\mathbf{g}_a^{(\text{QM})} - \mathbf{g}_a^{(\text{QM/MM})}|$. The dashed horizontal lines represent errors from pure MM calculations, which show that the AMOEBA parameters we obtained are highly accurate for the test systems. The bar charts represent the errors when QM/MM is used, and the bar colors are consistent with the corresponding QM region in Figure 1.

Single atom link scheme (SL): As already discussed in Table 2, the only parameter in the single link atom scheme (SL) is ω_{MM} , which controls how the QM region interacts with atoms in the MM region that are not MMHosts. Figure 2(a) shows the comparison of errors when different values of ω_{MM} are used in the SL scheme. The errors are notably affected by the choices of the QM region. When the QM/MM boundary cuts through C-C single bond between the backbone and the side chain (i.e. QM5, dark blue), the errors stay consistently small for both energies and gradients, and are insensitive to the choice of ω_{MM} . However, such QM/MM boundary is often infeasible for protein simulations, and cutting through the backbone is necessary in most situations. For all other choices of QM regions, the errors in the energies stay relatively large for small ω_{MM} (more delocalized charges), and becomes smaller for $\omega_{\text{MM}} \rightarrow \infty$ (i.e. treated as point charges). However, the behavior is quite different for gradients. In particular, as ω_{MM} becomes bigger and approaches point charges, we observe notable errors in gradients for QM2 (yellow) and QM6 (purple), both corresponding to cutting through C_α -C bond along the backbone. Examination of the gradient elements show that at $\omega_{\text{MM}} = \infty$, large errors in nuclear gradients are concentrated on atoms within 1 or 2 bonds from the QM/MM boundary, while errors on other MM atoms remain small. We think these large errors are directly related to the property that both the carbon atom (MMHost) and oxygen atoms of the carbonyl group have large partial charges. When such large charges are very close the QM region, it is no longer a good approximation to assume that the QM region sees the carbonyl group as point charges, and a more physical description can be obtained by treating them as delocalized charges when interacting with the QM region.

Default	MMLink/MMHost				Other MM
	KeepPerm	KeepInduce	Redistrib.	wLink (ω_{link})	wMM(ω_{MM})
SL	N/A				∞
DL(*)	Yes	No	Yes	0.3	∞

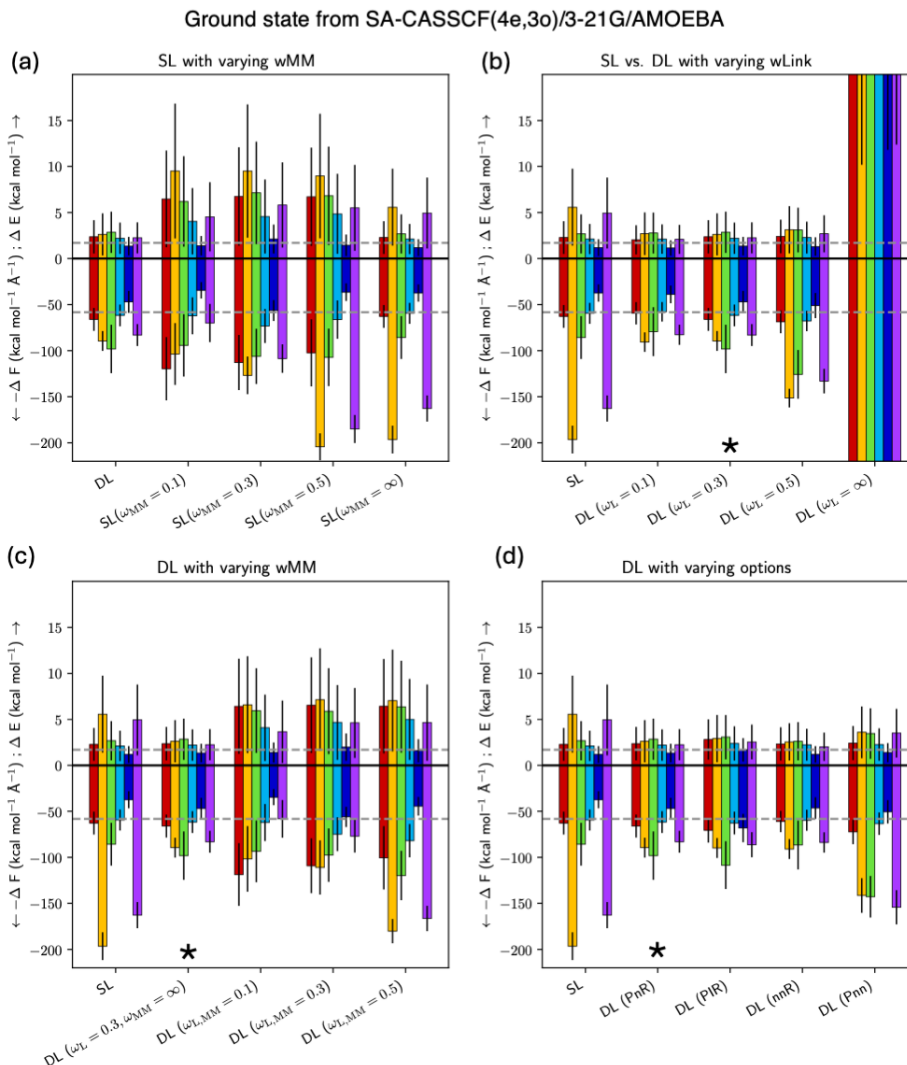


Figure 2: Statistics of errors of ground state properties computed for 1000 conformations of the molecule in Fig. 1 using QM/MM with different link atom schemes versus full QM calculations. The table at the top summarizes the default settings for single link atom (SL) and double link atom (DL) scheme respectively. Within each panel, colors of the bars represent different QM regions and are consistent with Fig.1. Positive bars correspond to mean absolute errors in relative potential energies and negative bars correspond to mean gradient errors as defined in the main text. Error bars represent one standard deviation. The dashed horizontal lines indicate errors from pure MM calculations. In panel (d), the three letters in the parenthesis specify choices of "KeepPerm-KeepInduce-Redistribute" respectively, where an uppercase letter means "yes" while "n" means "no". The * indicates the default setting of the DL scheme.

Based on these observations, although we choose $\omega_{\text{MM}} = \infty$ as the default value for SL, we are not able to find a value of ω_{MM} in the current SL scheme that can provide consistent reliable results for both energies and gradients. In particular, we have found that the largest errors are often correlated with the situation when MM atoms near the QM/MM boundary possess large point charges. These will be improved through the DL scheme as discussed below.

Double atom link scheme (DL): Double link atom scheme (DL) also contains the parameter ω_{MM} similar to the SL scheme. In addition, DL scheme offers a few more options related to the MMHost and MMLink atoms, which include: whether to keep the permanent dipoles and quadrupoles on the MMHost (KeepPerm), whether to keep the induced dipoles on the MMHost (KeepInduce), whether the permanent charges on the MMHost should be redistributed (Redistribute), and lastly the parameter ω_{link} that determines the interactions with the QM region.

We first study how the accuracy is affected by ω_{link} and ω_{MM} respectively. Figure 2(b) shows when ω_{MM} is fixed at ∞ , how the errors change with different values of ω_{link} . As shown in Figure 2(b), the errors in both energies and gradients remain small as ω_{link} changes from 0.1 to 0.3, while the errors in gradients become larger at $\omega_{\text{link}} = 0.5$ corresponding to more localized multipoles on MMHost and MMLink atoms. In the limit of point charges ($\omega_{\text{link}} = \infty$), very large errors are observed for both energies and gradients. We think these behaviors are again related to the facts that MMHost and MMLink atoms are too close to the QM region to be treated as point charges, and the resulting full Coulomb interaction likely causes overpolarization in the system. The small errors from $\omega_{\text{link}} = 0.1$ and 0.3 suggest that such overpolarization can be effectively mitigated with the damped Coulomb interaction, which is equivalent to delocalizing the charges on MMHost and MMLink atoms. Based on these observations, we choose $\omega_{\text{link}} = 0.3$ as the default value in DL scheme, which corresponds to roughly the vdW radius of the MMHost (length scale of $1/0.3$ a.u. = 1.76 Angstrom).

Calculations in Figure 2(b) delocalize the permanent multipoles on MMHost and MMLink atoms, while those on the remaining MM atoms are kept as point multipoles. The question remains whether the multipoles on the other MM atoms should also be delocalized, such that the entire MM region is treated in the same way. We test this in Figure 2(c) by comparing errors from such two tier treatment (i.e. $\omega_{\text{link}} = 0.3$, $\omega_{\text{MM}} = \infty$) with results from setting $\omega_{\text{link}} = \omega_{\text{MM}}$ to the same value. As shown in 2(c), except for QM5 (deep blue, cut through C-C single bond), $\omega_{\text{link}} = \omega_{\text{MM}}$ generally gives worse results. We think this is related to the fact that the AMOEBA force field are parametrized with the assumption that MM atoms are described with point multipoles, and the accuracy of the force field is degraded when the same parameters are used as delocalized multipoles in the QM/MM interaction term. Therefore, we think the multipoles on the remaining MM atoms should be kept as point multipoles, and we choose $\omega_{\text{MM}} = \infty$ as the default value in the DL scheme.

Figure 2 (d) tests various options regarding quantities on MMHost and MMLink atoms. The improvement from the redistribution algorithm on the MMHost permanent charges is obvious for QM boundary that cuts through C_{α} -C bond (QM2 yellow and QM6 purple) or the amide bond (QM3, green), where the errors without redistribution (Pnn) are notably larger than with redistribution (PnR). Thus, we choose to apply the redistribution algorithm as the default setting in DL scheme. Other than the redistribution algorithm, the errors are not very sensitive to whether the permanent multipoles and induced dipoles are turned on or off on MMHost/MMLink atoms. In SI-Figure S3, we have repeated similar tests as 2(b) and 2(c) but with the induced dipoles on MMHost/MMLink kept. As shown in SI-Figure S3, the trends with varying ω_{link} and ω_{MM} are similar with or without the induced dipoles. However, once the ω_{link} and ω_{MM} value exceed 0.5, keeping the induced dipoles often leads to larger errors, likely due to overpolarization. Based on these results, in the default setting of DL scheme, we set KeepPerm to true such that information from the AMOEBA force field is maximally preserved, and set KeepInduce to false to avoid potential overpolarization.

Testing of link-atom schemes on excited state properties

The results so far are obtained with the 3-21G basis set, and have only been applied to ground state properties. We now need to test if the link atom scheme is applicable to excited state properties.

Using the default settings for SL and DL schemes, in Figure 3(a), we have repeated the same test as Figure 2 for ground state properties but using the cc-pVDZ basis set, which has additional p functions on hydrogen atoms and additional d functions on all other atoms. The errors from the pure MM calculations (horizontal dashed lines) become bigger, especially that the gradient errors almost doubled. This is expected since the force field parameters were fitted to calculations using 3-21G. The errors from QM/MM calculations don't change as much, and the results from cc-pVDZ behave quite consistently with the results from 3-21G. In particular, we see that DL scheme is more advantageous than the SL scheme in that the errors are less sensitive to the choices of the QM regions. In addition, we have repeated the same test but for the first excited state properties in Figure 3(b). The pure MM calculations are qualitatively wrong in this case since the parameters are only designed to reproduce the ground state properties. When DL scheme is used, the errors are quite consistent between ground state and excited state. In contrast, when SL scheme is used, we observe a notable increase in errors for the excited state when QM Region 3 is used (green).

To help understand the discrepancy between the two link atom schemes for excited states, in Figure 3, we present the same data from a different perspective and focus on analyzing the first excitation energy instead. In Figure 3(c), we use box plots to present the distribution of errors in the first excitation energy, i.e. $\Delta E_{S1-S0}^{(QM)} - \Delta E_{S1-S0}^{(QM/MM)}$. When DL scheme is used, the mean absolute error in the first excitation energies is 0.020 eV for QM region 2 and smaller for other QM regions, and all are within 0.1eV for all sampled structures and for all QM regions. In contrast, when SL scheme is used, the mean absolute error for QM region 3 (green) is 0.264 eV and we observe several outliers where the errors exceed 1eV. The errors in QM region 3 using SL are significantly larger than the results from DL, which is consistent with what we observed in 3(b). The notable difference between SL and DL for

QM region 3 is further presented in 3(d), which shows the scatter plot that compares the QM/MM excitation energies with full QM results. The excitation energies from DL scheme (black dots) are concentrated on the diagonal line, indicating good agreement with full QM results. In contrast, the SL data are separated into two groups, with one group close to the diagonal line, while the other group notably underestimating the first excitation energy.

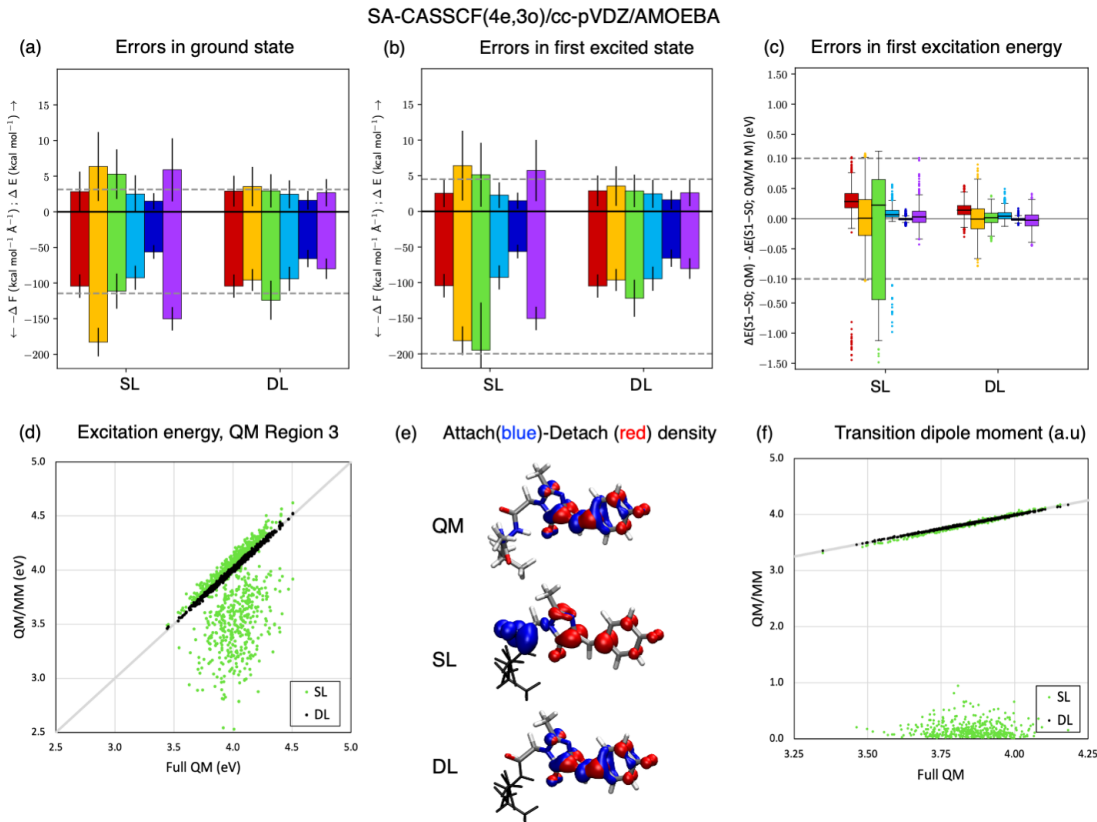


Figure 3: Statistics of errors from computing the (a) S0 and (b) S1 properties using QM/MM versus full QM for the same 1000 conformations as in Fig. 2. Both SL and DL use the same default settings as in Fig.2. (c) Errors in the excitation energies presented as box plots using Tukey’s definition where the box ranges from 25-75 percentile, whiskers are $1.5 \times$ the interquartile range and points indicate outliers.⁶⁵ The y-axis for errors greater than ± 0.1 eV is scaled for improved readability. (d) Scatter plots of QM/MM excitation energies using QM region 3 versus full QM, which has the largest difference between SL (green) and DL (black). For the structure that has the largest error in (d) using SL, (e) shows the attachment-detachment density analysis for full QM, SL and DL respectively. (f) shows the scatter plots of QM/MM transition dipole moments versus full QM results corresponding to QM region 3. Both attach-detach analysis in (e) and transition dipole moments in (f) can be considered as indicators for the characteristics of the excited state.

In order to understand the source of these large errors, we took the structure where SL produced the largest error and applied attachment-detachment analysis to analyze the characteristics of the corresponding excitation. As shown in Figure 3(e), the result from the DL scheme agrees well to the full QM calculations, and both involve excitations on the chromophore. In contrast, the result from the SL scheme is qualitatively different, and predicts an excitation from the chromophore to atoms near the QM/MM boundary. We think this is probably related to how DL and SL treat the MMHost differently. The SL scheme zeros out a significant negative charge on the MMHost (-0.6868) and distributes it evenly on the other MM atoms; this significantly increases the electrostatic potential close to the MMHost and lowers the energy for electrons to be excited into this region. On the other hand, the DL scheme transfers part of the negative charge on the MMHost to positively charged atoms bonded to the MMHost, and this does not perturb the MM electrostatic potential as greatly as the SL scheme. Along similar lines, the DL scheme has a partial negative charge on the MMLink that has a repulsive effect on electrons excited into this region, whereas this charge is distributed throughout the MM region in the SL scheme.

Because the characteristics of the excited state is also reflected in the transition dipole moment, we show the scatter plot that compares the transition dipole moments from QM/MM to full QM calculations in Figure 3(f). The results from DL scheme stay close to the diagonal line, indicating that the small errors in its excitation energy are correlated with being able to find the correct excitation states. In contrast, the transition dipole moments from the SL scheme are clearly separated into two groups, with one close to the diagonal lines, while the other group is much lower in the transition dipole moment. These results show that the large errors in the SL scheme are directly related to the situations when an incorrect excitation state is identified. Although a large QM region can prevent this from happening, this may not always be computationally feasible, and the choice of the link atom scheme will be important for providing the correct electrostatic interaction near the QM/MM boundary.

3.2 Application to NanoLuc

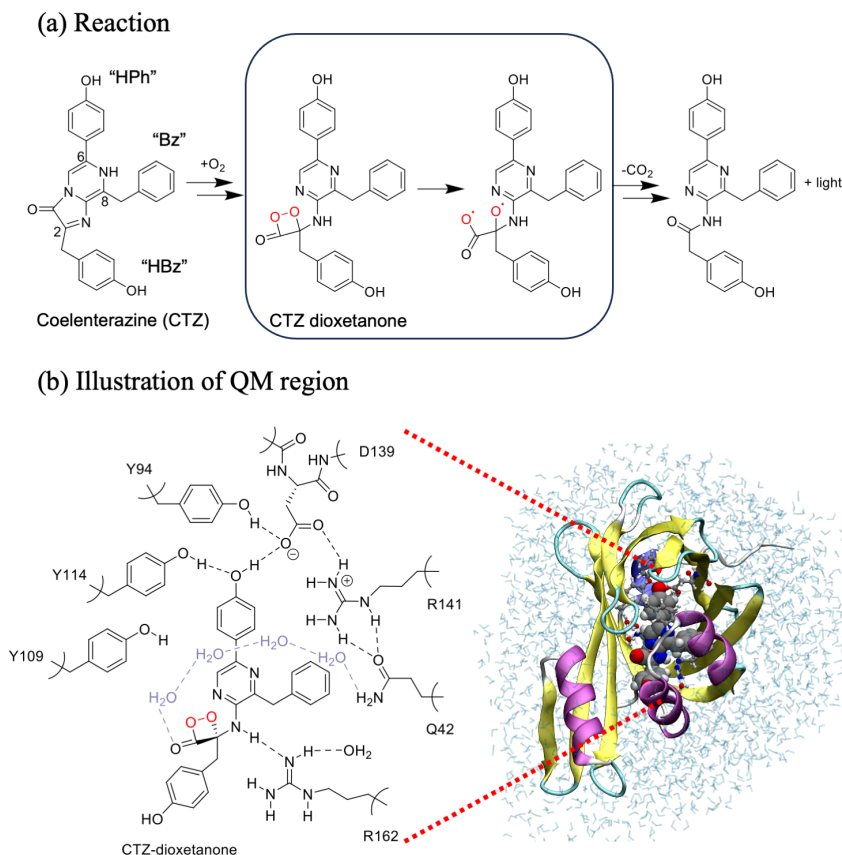


Figure 4: (a) Bioluminescence reaction of coelenterazine (CTZ). The box highlights the elementary reaction studied in this work, which involves the dissociation of the red O-O bond in CTZ dioxetanone. (b) Illustration of QM/MM models of NanoLuc. The chemical structure on the left shows what side chains and water molecules are included in the QM region in addition to the chromophore. In the 3D structure on the right, the chromophore is rendered as vdW spheres. Other atoms in the QM region are rendered using CPK (ball and stick). MM atoms in the protein are rendered according to their secondary structure. MM atoms in solvent water are rendered as lines.

To test the method on real reactions in protein environment, we applied SA-CASSCF/AMOEBA to the studies of an elementary reaction step in NanoLuc,⁶⁶ which is an engineered luciferase that catalyzes bioluminescent reaction.

The chromophore in NanoLuc is coelenterazine (CTZ), and its structure is shown in Figure 4(a). Its chemical structure contains imidazopyrazine (will be referred to as the “core”) in which positions 2, 6, and 8 are substituted by 4-hydroxybenzyl (“HBz”), 4-hydroxyphenyl

(“HPh”), and benzyl groups (“Bz”), respectively. As described in a mechanistic study by Nemergut et al,⁶⁷ the bioluminescence process is initiated by CTZ reacting with an oxygen molecule. Through a series of elementary reactions, CTZ is eventually transformed into CTZ dioxetanone (see Figure 4(a)). The red O-O bond in the CTZ dioxetanone then dissociates, followed by the leaving of CO₂. The hypothesis is that the leaving of CO₂ provides the thermal activation energy needed to promote the system to the excited state, which makes light emission possible.

Due to the complexity of the bioluminescence mechanism, we chose not to simulate the entire reaction pathway in this work. Instead, we will mainly focus on the elementary reaction step of O-O bond dissociation of CTZ dioxetanone as shown in the box of Figure 4(a). Despite that this reaction step primarily proceeds on the ground state, it still makes a good test for the newly developed SA-CASSCF/AMOEBA for the following reasons. First, a multi-reference approach is required for describing the prominent strong correlation that arises during the bond dissociation process. Second, past theoretical studies on dioxetanone in the gas phase have suggested as the O-O bond dissociates, the energy gap between ground state and $^1(n, \sigma^*)$ excited state gradually approaches zero, and this property may play an important role to facilitate transfer of molecules to the excited state for light emission.^{68,69} As a result, it will be important to understand the behaviors of both ground and excited states along the reaction pathway.

Below we will first discuss how the QM/MM system is set up for SA-CASSCF/AMOEBA calculations, and we will then discuss the results that show how the reaction mechanisms differ in gas phase, PCM, and QM/MM system. Geometry optimizations (including unconstrained or constrained minimization, and minimal energy conical intersection search) are performed using geomeTRIC software.^{70,71} Interpolation pathways are generated using linear interpolation of the translation-rotation internal coordinates (TRIC), a delocalized internal coordinate system that explicitly treats fragment translation and rotation degrees of freedom.

Setup of QM/MM system

We start with the PDB structure 8BO9⁶⁷ corresponding to NanoLuc complexed with aza-CTZ, a non-oxidizable analogue of CTZ. The PDB file contains three chains of similar structures, and we only keep the chain that has the greatest number of crystallographic waters for our studies. In particular, the chain we kept contains one crystallographic water near the chromophore. The coordinates of aza-CTZ are converted to CTZ by substituting N with C at position 2 of the core, and the HPh group is assumed to be deprotonated, thus setting the charge on CTZ to -1 . The nearest water molecule to CTZ is replaced with O₂. H++ software⁷² is used to determine protonation states of residues and adding hydrogens correspondingly, and PDBFixer⁵⁷ is used to add solvation of a rectangular box with a 1.0 nm water padding around the protein. To obtain force field parameters for CTZ, the geometry is optimized using B3LYP-D3/6-31G*, then the electrostatic parameters are obtained by fitting the electrostatic potential of MP2/6-311G(d,p) following the recommended procedure.²⁹ The AMOEBA parameters for CTZ are provided in the SI. For the O₂ molecule, the permanent multipoles are all set to zero, the bond length is set to the gas phase experimental value of 1.207 Å, while other parameters are borrowed from carbonyl oxygen parameters in the AMOEBA organic molecule force field.²⁹ AMOEBA protein force field parameters are used for amino acids,²⁷ and the AMOEBA14 water model is used for solvents.²⁵ The solvated complex is then equilibrated by 10 ns of classical MD simulation using the AMOEBA force field under periodic boundary conditions. Because we did not optimize the intramolecular force field parameters for CTZ, its conformation was restrained by the addition of a 100 kJ/mol/nm² harmonic restraint force applied to the Cartesian positions of each atom. At the end of the equilibration simulation, the structure of the solvated system is used for further QM/MM calculations.

To create initial structure for any QM/MM geometry optimizations, two modifications are made on the equilibrated snapshot. First, only water molecules within 5 Å from the protein or within 20 Å from the chromophore are kept, which reduced the simulation size

from 24,219 to 8,609 atoms. Second, the CTZ+O₂ coordinates in the snapshot is manually modified to the desired conformation (i.e. dioxetanone in the reactant) before QM/MM optimizations, as described in details in SI-Section 5. Here we choose not to rerun the MM equilibration after the substitution of dioxetanone. This is because the dioxetanone is a reactive intermediate, thus its lifetime is unlikely to be sufficiently long to allow the entire protein and solvents to fully equilibrate with respect to its structure. Instead, we choose to only relax atoms within a selected region surrounding the CTZ dioxetanone using QM/MM geometry optimizations as described below. Note that there are two different regions during QM/MM geometry optimization. One is the QM region as illustrated in Figure 4(b). The QM region contains 177 atoms, which includes the chromophore along with several nearby side chains that form hydrogen bond with the chromophore. A second region is the “relaxed region” as illustrated in SI-Figure S4. As the name implies, the positions of atoms within the relaxed region will be relaxed during geometry optimizations or interpolations, while atoms outside the relaxed region will stay fixed at the positions from the snapshot. The relaxed region contains 932 atoms, which includes the entire QM region, as well as several hydrogen-bonding, polar, nonpolar residues, and water molecules surrounding the QM region. The relaxed region was chosen by selecting all atoms within around 5 Å of the chromophore and extending it to include complete residues and joining together segments that are separated by 1-2 residues.

In all the QM/MM calculations, AMOEBA force field parameters for the MM atoms are identical to the ones used in the pure MM equilibration simulation. The method for the QM region is SA-CASSCF(8e,9o) using 6-31G* basis sets, which includes sufficient active orbitals for describing the O-O bond dissociation of interest. Four singlet states and two triplet states are included in the state-averaging in the calculations. Because the reaction proceeds on the ground state, S₀ is chosen as the reference state for determining the dynamic weights of states using a bandwidth of 3 eV. For comparison, we have also included calculations in gas phase and PCM($\epsilon = 78.4$) at the same level of theory.

Optimized structures of reactant and product using QM/MM

The MM equilibration simulation discussed in the previous paragraph was designed to be consistent with the hypothesis in Nemergrut et.al,⁶⁷ who proposed that CTZ is deprotonated at HPh throughout the reaction by D139 and is protonated at N1 by R162, and is therefore an anion overall. This hypothesis is based on experimental observation of a redshift in the CTZ absorption when in complex with a *Renilla* luciferase mutant (RLuc8), and the disappearance of the redshift following a D162A (aspartic acid \rightarrow alanine) mutation.⁷³ After replacing the CTZ + O₂ with a model of the CTZ dioxetanone and carrying out QM/MM unconstrained geometry optimizations of this intermediate, we found that HPh is always protonated and the D139 is deprotonated in the optimized structure regardless of which group is protonated initially. Optimization of the deprotonated state can only be achieved by applying constraints on the O-H distances in the proton transfer region and increases the energy relative to the HPh-protonated structure by 15.6kcal/mol. Based on these results, we decided to keep the HPh group protonated and D139 deprotonated throughout our studies. However, it is worth noting that because SA-CASSCF lacks dynamic correlation, the method is inherently not good at describing the dispersion interactions that contribute to hydrogen bonding.^{74,75} Therefore, this problem should be revisited in the future when we have MS-CASPT2 to capture dynamic correlation on top of SA-CASSCF.

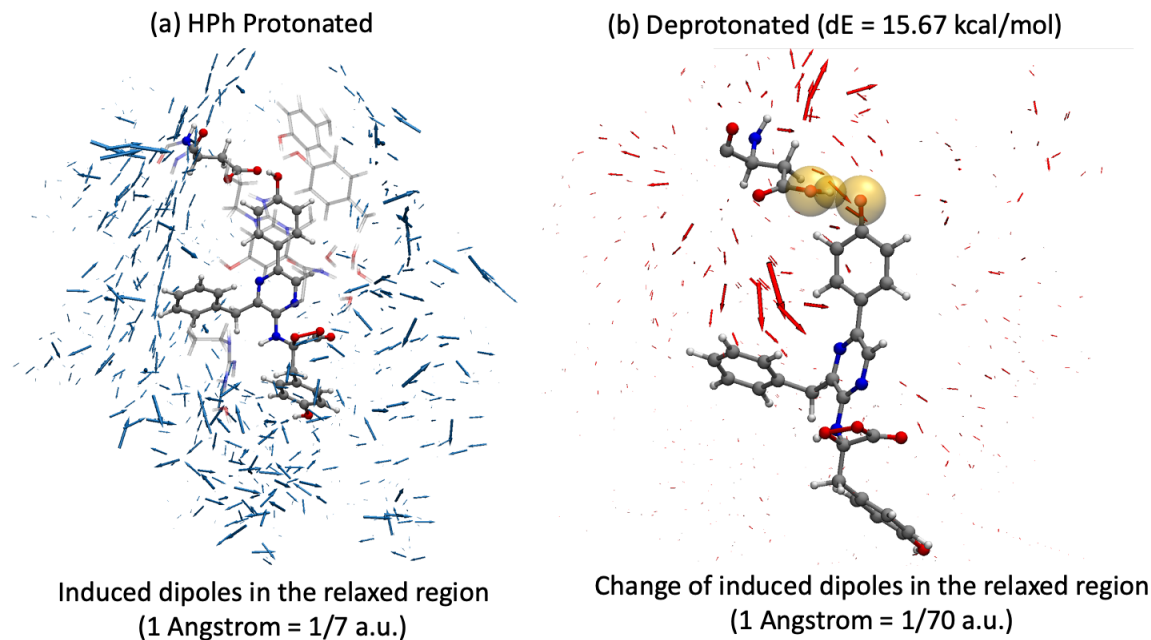


Figure 5: (a) Induced dipoles in the relaxed region corresponding to the optimized reactant structure, where the HPh group on the CTZ dioxetanone is protonated. (b) Changes of the induced dipoles in the relaxed region after the proton is transferred from HPh group to D139. The energy of the deprotonated structure is minimized through constrained optimization.

The induced dipoles in the AMOEBA model should spontaneously respond to any change in the QM region. As an additional verification of the method, Figure 5(a) shows the induced dipoles surrounding the QM region when the HPh group in CTZ dioxetanone is protonated, and Figure 5(b) examines the change in these MM induced dipoles once the proton moves from CTZ dioxetanone to the nearby D139. As shown by the lengths and directions of the red arrows, the induced dipoles mainly change near the proton transfer region, while those away from the proton transfer largely stay the same. In addition, after the proton transfer, we expect that the donating oxygen becomes more negative in charge while the accepting oxygen becomes more positive in charge. In respond to this, we observed that the induced dipoles change to point towards the donating oxygen on CTZ dioxetanone and point away from the accepting oxygen on D139. These observations are consistent with the expected behavior of the induced dipoles, and provide further evidence that the method is implemented correctly. In SI-Figure S5, we have also compared the above proton transfer

energy with SA-CASSCF(8e,9o)/6-31G* calculations in fixed charge force field, which was set up in AMBER using the AMBER-FB15 protein force field,⁶⁴ SPC/Fw flexible water model,⁷⁶ and GAFF forcefield⁷⁷ for CTZ dioxetanone. The geometries of both protonation states in the relaxed region were reoptimized using SA-CASSCF in fixed charge model. To investigate the dependence of the proton transfer energy on the QM region size, we further carried out single point calculations on the optimized structures using four different QM region sizes that are different subsets of the original QM region. As shown in SI-Figure S5, the proton transfer energy computed using the AMOEBA QM/MM model is more stable with respect to the QM region size compared to the fixed-charge model, especially in the case of the smallest QM region where the QM/MM boundary goes through the chromophore. One possible explanation is that the AMOEBA force field more closely reproduces the response of the QM system to the structural changes, in particular the change in polarization response caused by the proton transfer. This indicates that future QM/MM studies with polarizable embedding could benefit from using smaller QM regions, although more extensive benchmark studies are needed to confirm this.

We have also carried out QM/MM geometry optimization of the product structure in NanoLuc. As shown in Figure 6 (c), the positions of the chromophore side-groups before and after the O-O bond dissociation stay almost the same in the protein environment. In contrast, when optimizing only the chromophore structure either in the gas phase (Figure 6(a)) or in PCM (Figure 6(b)), there are notable changes in the positions of HBz group when comparing the reactant and product. This different behavior of QM/MM versus gas phase or PCM is likely related to the strong steric effect in the cavity of protein, which restrains the motions of the chromophore during the reaction. In SI-Figure S6, we have also presented the same results from a different perspective by showing how the reactant and product structures differ in gas phase, PCM and QM/MM. For both reactant and product, the structures in the gas phase and PCM are quite similar except for a small rotation in the HBz group. The QM/MM structure is much more twisted likely due to steric effect again. For example,

the large twist in the Bz group in the protein compared to in the gas phase is likely to avoid collision with the hydrophobic side chains of L18 and L22. These comparisons show that steric effects are important for predicting the geometry of the chromophore in protein environment, and using atomistic models is important for properly capturing such effects.

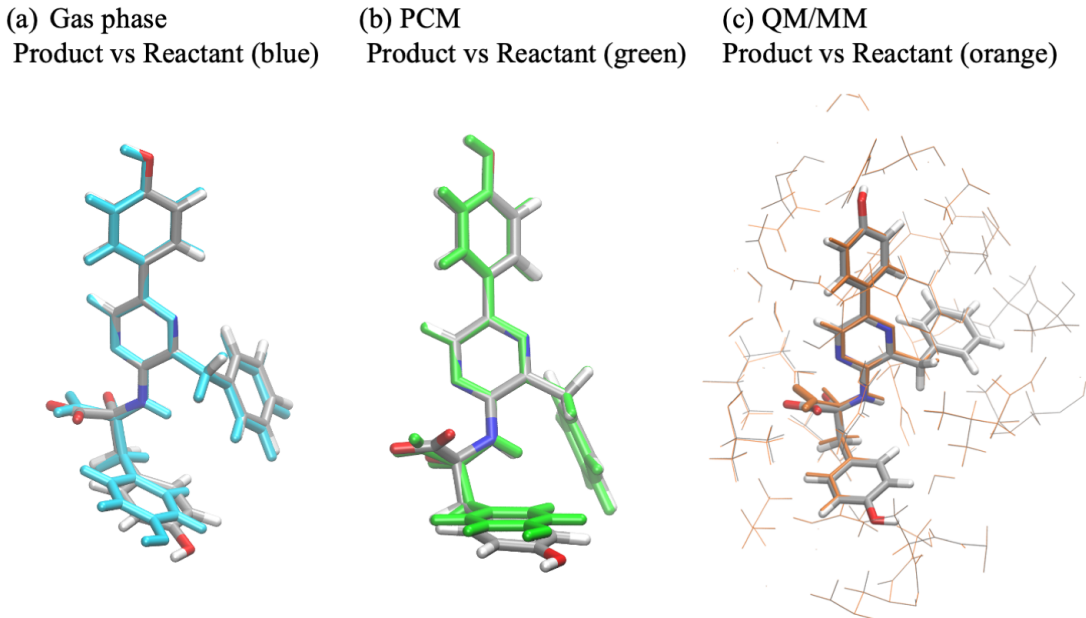


Figure 6: Comparison between optimized structures of the reactant and product using SA-CASSCF(8e,9o)/6-31G* in (a) gas phase, (b) PCM ($\epsilon = 78.4$) and (c) QM/MM respectively.

Potential energy curves along reaction pathways

Given the optimized reactant and product structures, we then generate an interpolation path connecting the reactant and product and compute the potential energy curves along the pathway. The results are presented in Figure 7.

The gas phase potential energy curves are shown in Figure 7(a). There are a few intersections with higher electronic states during the initial part of the pathway, but the characteristics of the states stay unchanged as the reaction proceeds. To understand the characteristics of each electronic state, we have carried out attach-detach density analysis at image 20 along the path. From the analysis results presented in SI-Figure S7, S1 through S3 are all local excitations $^1(n, \sigma^*)$ in the O-O bond region. The two triplet states are also

local excitations, except that T1 is $^3(n, \sigma^*)$ while T2 is $^3(\sigma, \sigma^*)$. One prominent feature of the gas phase results is that the energy gap between S1 and S0 gradually closes as the O-O bond dissociates, which is consistent with previous theoretical studies on dioxetanone in the gas phase.^{68,69}

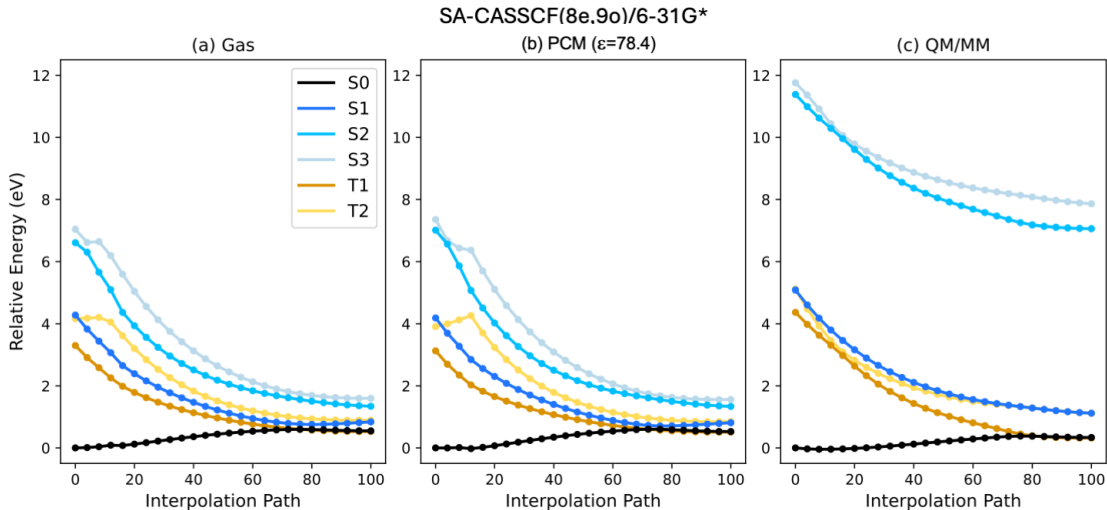


Figure 7: Potential energy curves along the interpolation path from reactant to product in (a) gas phase, (b) PCM and (c) QM/MM respectively. Geometries along the path are generated by interpolating the optimized reactant and product structures shown in Fig.6 for each environment separately. Dynamic weights are applied in the PCM and QM/MM calculations, where the ground state is set as the reference state with a bandwidth of 3eV.

The potential energy curves in the PCM environment are presented in Figure 7(b), and are quite similar to the gas phase results. The attach-detach analysis in SI-Figure S7 shows that excitations in PCM share similar characteristics from the gas phase calculations, and the S0 and S1 also become energetically close as the O-O bond dissociates. The similarity between PCM and gas phase is likely related to the fact that these low lying electronic states mainly involve local excitations, thus the electrostatic response from the solvent do not change significantly between the ground and excited states. Therefore, the excitation energies are not notably affected when PCM is used as the solvent model.

The potential energy curves in QM/MM (Figure 7(c)) are quite different from either gas phase or PCM results. One major difference is that S2 and S3 become much higher in

energy in QM/MM. This may be due to that S2 and S3 in QM/MM also involve excitations in the core region, as revealed by the attach-detach analysis in SI-Figure 7. Another major difference is that the energy gap between S0 and S1 in QM/MM no longer approaches zero as the O-O bond dissociates. The intersection between S0 and S1 is further investigated using conical intersection minimization as discussed below. These observations all suggest that the potential energy curves can be qualitatively modified by the protein environment, which should be taken into account for accurate simulation of reaction mechanism.

Search for minimal energy conical intersection (MECI):

To further understand the behavior of how the energy gap between S0 and S1 closes, we search for minimal energy conical intersections in the vicinity of the product region. Because in gas phase and PCM, the energy gap is already very close to zero as the bond dissociates, the corresponding MECI geometry closely resembles the product structure as shown by the comparison in SI-Figure S8. Nevertheless, despite their similarity in the MECI geometry, the shapes of the potential energy surfaces around the conical intersection are still quite different between gas phase (Figure 8(a)) and PCM (Figure 8(b)). This shows that the shapes of conical intersections can be quite sensitive to the electrostatic effects from the environments.

The MECI in QM/MM is quite different both in terms of its molecular geometry as well as the topology around it. In terms of the molecular geometry, the structure of MECI in QM/MM no longer resembles the product, which is consistent with our previous observation that energy gap of S0/S1 does not close along the reaction pathway. Instead, as shown in Figure 8(d), the O-C-O group in the QM/MM system needs to rotate by almost 90 degrees from the original plane of the dioxetanone 4-membered ring in order for S0 and S1 to come close and reach conical intersection. In terms of the topology, the MECI in PCM is peaked, as the red contour lines in Figure 8(b) indicate that its S1 local minimum also occurs at the conical intersection. In contrast, the MECI in QM/MM is sloped, as the red contour lines in Figure 8(c) suggest that S1 local minimum is no longer at the MECI but is located along

the positive g -vector direction. Theoretical studies have suggested that whether the MECI is peaked or sloped may affect the dynamics of population transfer between the two states, which provides another way for the protein environment to affect the reaction mechanism.⁷⁸

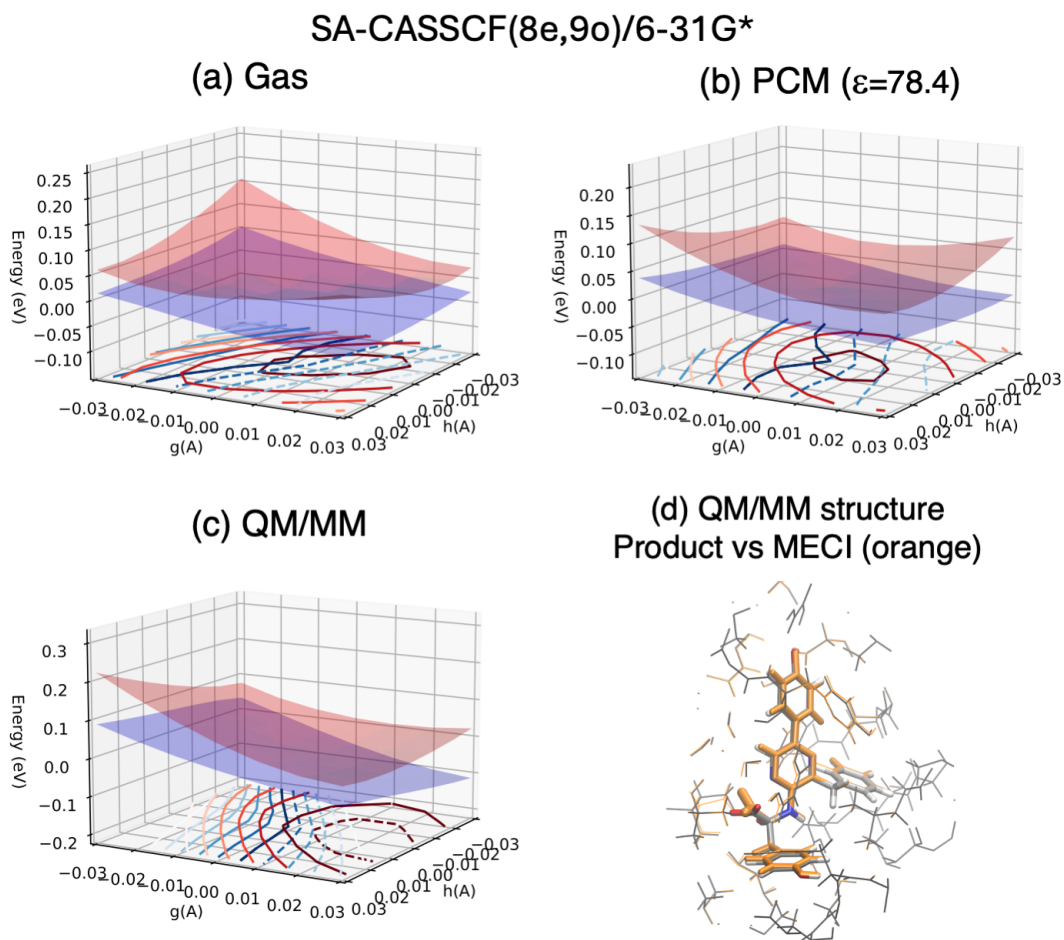


Figure 8: Potential energy surfaces around the minimal energy conical intersections (MECI) for (a) gas phase, (b) PCM, and (c) QM/MM. In each panel, the red and blue contours lines correspond to the isoenergy curves of S1 and S0 respectively. (d) compares the MECI structure with the optimized product structure in QM/MM.

In summary, SA-CASSCF/AMOEBA simulations on the O-O bond dissociation of CTZ dioxatone in NanoLuc lead to qualitatively different results from simulations in the gas phase and PCM. Through interactions with the protein environment, we have observed changes in the optimized structures, characteristics of the excited states, and properties of the conical intersections.

4 Conclusions

In this work, we have developed two important ingredients needed to make SA-CASSCF/AMOEBA feasible for studying photoreactions in protein environments, i.e. intramolecular polarization and the link atom scheme. We have shown that the inconsistency between direct and polarization electric fields can be naturally incorporated into the SA-CASSCF framework with very little additional computational cost. We have also shown that the double link atom scheme can provide consistent accurate results for both ground and excited state properties.

The success of the this work paves the way towards simulating photo-reactions in protein with rigorous descriptions of the protein environmental effects. There are a few improvements that can be further addressed in future works. Recently, improvements to the AMOEBA functional form have been described including charge penetration, charge transfer, and geometry-dependent charge flux terms, and these have been implemented into the AMOEBA+⁷⁹ and HIPPO⁸⁰ models. Many of these improvements were modeled after the short-range behavior of the QM density, and therefore these force fields may have improved compatibility with QM methods, though a careful treatment of the QM/MM boundary would still be needed.

In terms of the QM method, as mentioned in the discussions about NanoLuc, the SA-CASSCF method used in this work lacks dynamic correlation needed for dispersion interactions, thus cannot accurately describe the hydrogen bonding interaction within the QM region. The simplest way is to capture the dynamic correlation through second order perturbation theory, and the Fock operator from SA-CASSCF/AMOEBA can be used as the zeroth order Hamiltonian for extended multi-state CASPT2 (XMS-CASPT2) calculations.⁸¹⁻⁸³ In particular, a dynamic weight scheme very similar to what we used in this work has recently been incorporated into the framework of XMS-CASPT2,⁸⁴ with similar purpose of interpolating between state-specific Fock operator at the Franck-Condon region and the equally state averaged Fock operator at the conical intersection region. As an initial step for XMS-CASPT2/AMOEBA calculations, we will only include the effects from AMOEBA in the

zeroth order Hamiltonian, and we will determine whether this is sufficient by carrying out benchmark studies on the errors from XMS-CASPT2/AMOEBA calculations versus full QM XMS-CASPT2 calculations.

Another future improvement needed is more efficient treatment of the MM induced dipoles. The NanoLuc studied in this work is a small protein (19.1 kDa). After including the solvent water molecules, the entire MM region is within $\sim 10,000$ atoms, and solving the corresponding MM induced dipoles is not a computational bottleneck so far. However, many interesting photoreactions take place in much larger protein (e.g. molecular weight of phytochrome is approximately 120 kDa), and the corresponding solvent water sphere will also be much bigger. Such large protein is currently beyond the capabilities of our code, because interactions involving the MM induced dipoles are treated in the same way regardless of their distance. This results in a quadratic scaling increase in the computational cost with respect to the number of MM atoms for our current implementation. The fast multipole method has been shown to be effective in reducing the computational cost of QM/AMOEBA to linear scaling,⁸⁵ and we plan to incorporate this idea into SA-CASSCF/AMOEBA in future works.

Acknowledgements

C.S acknowledges the support from the startup funds provided by UC Davis chemistry department. L.-P.W. is grateful for a restricted gift to the UC Davis Chemistry Department from Relay Therapeutics, Inc.

Associated Content

Supporting Information Available:

Details about the redistribution algorithm of MMHost permanent charges in double link atom scheme, procedure to generate AMOEBA parameters for the test systems, as well as tables and figures that present additional test results.

The Supporting Information is available free of charge on the ACS Publications website.

References

- (1) Björn, L. O. *Photobiology: The Science of Life and Light*; Springer New York, NY, 2008.
- (2) Solovchenko, A. *Photoprotection in Plants: Optical Screening-based Mechanisms*; Springer Berlin, Heidelberg, 2010.
- (3) Burgie, E. S.; Vierstra, R. D. Phytochromes: An Atomic Perspective on Photoactivation and Signaling. *The Plant Cell* **2014**, *26*, 4568–4583.
- (4) Hofmann, K. P.; Lamb, T. D. Rhodopsin, Light-Sensor of Vision. *Progress in Retinal and Eye Research* **2023**, *93*, 101116.
- (5) Fleiss, A.; Sarkisyan, K. S. A Brief Review of Bioluminescent Systems (2019). *Current Genetics* **2019**, *65*, 877–882.
- (6) Wang, Y.; Han, N.; Li, X.; Yu, S.; Xing, L. Artificial Light-Harvesting Systems and Their Applications in Photocatalysis and Cell Labeling. *ChemPhysMater* **2022**, *1*, 281–293.
- (7) Christou, N.-E.; Apostolopoulou, V.; Melo, D. V. M.; Ruppert, M.; Fadini, A.; Henkel, A.; Sprenger, J.; Oberthuer, D.; Günther, S.; Pateras, A.; Rahmani Mashhour, A.; Yefanov, O. M.; Galchenkova, M.; Reinke, P. Y. A.; Kremling, V.; Scheer, T. E. S.; Lange, E. R.; Middendorf, P.; Schubert, R.; De Zitter, E.; Lumbao-Conradson, K.; Herrmann, J.; Rahighi, S.; Kunavar, A.; Beale, E. V.; Beale, J. H.; Cirelli, C.; Johnson, P. J. M.; Dworkowski, F.; Ozerov, D.; Bertrand, Q.; Wranik, M.; Bacellar, C.; Bajt, S.; Wakatsuki, S.; Sellberg, J. A.; Huse, N.; Turk, D.; Chapman, H. N.; Lane, T. J. Time-Resolved Crystallography Captures Light-Driven DNA Repair. *Science* **2023**, *382*, 1015–1020.

- (8) Pandey, S.; Bean, R.; Sato, T.; Poudyal, I.; Bielecki, J.; Cruz Villarreal, J.; Yefanov, O.; Mariani, V.; White, T. A.; Kupitz, C.; Hunter, M.; Abdellatif, M. H.; Bajt, S.; Bondar, V.; Echelmeier, A.; Doppler, D.; Emons, M.; Frank, M.; Fromme, R.; Gevorkov, Y.; Giovanetti, G.; Jiang, M.; Kim, D.; Kim, Y.; Kirkwood, H.; Klimovskaia, A.; Knoska, J.; Koua, F. H. M.; Letrun, R.; Lisova, S.; Maia, L.; Mazalova, V.; Meza, D.; Michelat, T.; Ourmazd, A.; Palmer, G.; Ramilli, M.; Schubert, R.; Schwander, P.; Silenzi, A.; Sztuk-Dambietz, J.; Tolstikova, A.; Chapman, H. N.; Ros, A.; Barty, A.; Fromme, P.; Mancuso, A. P.; Schmidt, M. Time-Resolved Serial Femtosecond Crystallography at the European XFEL. *Nature Methods* **2020**, *17*, 73–78.
- (9) Lin, H.; Truhlar, D. G. QM/MM: What Have We Learned, Where Are We, and Where Do We Go from Here? *Theoretical Chemistry Accounts* **2007**, *117*, 185–199.
- (10) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.
- (11) Park, J. W.; Al-Saadon, R.; MacLeod, M. K.; Shiozaki, T.; Vlaisavljevich, B. Multireference Electron Correlation Methods: Journeys along Potential Energy Surfaces. *Chemical Reviews* **2020**, *120*, 5878–5909.
- (12) Roos, B. O.; Lindh, R.; Malmqvist, P.; Veryazov, V.; Widmark, P.-O. *Multiconfigurational Quantum Chemistry*; John Wiley & Sons, 2016.
- (13) Matsika, S. Electronic Structure Methods for the Description of Nonadiabatic Effects and Conical Intersections. *Chemical Reviews* **2021**, *121*, 9407–9449.
- (14) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. M. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-CI Approach. *Chemical Physics* **1980**, *48*, 157–173.
- (15) Roos, B. O. The Complete Active Space SCF Method in a Fock-Matrix-Based Super-CI Formulation. *International Journal of Quantum Chemistry* **1980**, *18*, 175–189.

- (16) Schmidt, M. W.; Gordon, M. S. The Construction and Interpretation of Mcscf Wavefunctions. *Annual Review of Physical Chemistry* **1998**, *49*, 233–266.
- (17) Curchod, B. F. E.; Martínez, T. J. Ab Initio Nonadiabatic Quantum Molecular Dynamics. *Chemical Reviews* **2018**, *118*, 3305–3336.
- (18) Nelson, T. R.; White, A. J.; Bjorgaard, J. A.; Sifain, A. E.; Zhang, Y.; Nebgen, B.; Fernandez-Alberti, S.; Mozyrsky, D.; Roitberg, A. E.; Tretiak, S. Non-Adiabatic Excited-State Molecular Dynamics: Theory and Applications for Modeling Photo-physics in Extended Molecular Materials. *Chemical Reviews* **2020**, *120*, 2215–2287.
- (19) Pulay, P. A Perspective on the CASPT2 Method. *International Journal of Quantum Chemistry* **2011**, *111*, 3273–3279.
- (20) Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. Multiconfiguration Self-Consistent Field and Multireference Configuration Interaction Methods and Applications. *Chemical Reviews* **2012**, *112*, 108–181.
- (21) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable Embedding QM/MM: The Future Gold Standard for Complex (Bio)Systems? *Physical Chemistry Chemical Physics* **2020**, *22*, 14433–14448.
- (22) Nottoli, M.; Cupellini, L.; Lipparini, F.; Granucci, G.; Mennucci, B. Multiscale Models for Light-Driven Processes. *Annual Review of Physical Chemistry* **2021**, *72*, 489–513.
- (23) Saltiel, J.; D’Agostino, J. T. Separation of Viscosity and Temperature Effects on the Singlet Pathway to Stilbene Photoisomerization. *Journal of the American Chemical Society* **1972**, *94*, 6445–6456.
- (24) Vong, A.; Widmer, D. R.; Schwartz, B. J. Nonequilibrium Solvent Effects during Photodissociation in Liquids: Dynamical Energy Surfaces, Caging, and Chemical Identity. *The Journal of Physical Chemistry Letters* **2020**, *11*, 9230–9238.

- (25) Laury, M. L.; Wang, L.-P.; Pande, V. S.; Head-Gordon, T.; Ponder, J. W. Revised Parameters for the AMOEBA Polarizable Atomic Multipole Water Model. *The Journal of Physical Chemistry B* **2015**, *119*, 9423–9437.
- (26) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **2003**, *107*, 5933–5947.
- (27) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation* **2013**, *9*, 4046–4063.
- (28) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *Journal of Chemical Theory and Computation* **2018**, *14*, 2084–2108.
- (29) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *Journal of Chemical Theory and Computation* **2011**, *7*, 3143–3161.
- (30) Simmonett, A. C.; Pickard, F. C., IV; Shao, Y.; Cheatham, T. E., III; Brooks, B. R. Efficient Treatment of Induced Dipoles. *The Journal of Chemical Physics* **2015**, *143*, 074115.
- (31) Walker, B.; Liu, C.; Wait, E.; Ren, P. Automation of AMOEBA Polarizable Force Field for Small Molecules: Poltype 2. *Journal of Computational Chemistry* **2022**, *43*, 1530–1542.
- (32) Loco, D.; Lagardère, L.; Caprasecca, S.; Lipparini, F.; Mennucci, B.; Piquemal, J.-P. Hybrid QM/MM Molecular Dynamics with AMOEBA Polarizable Embedding. *Journal of Chemical Theory and Computation* **2017**, *13*, 4025–4033.

- (33) Dziedzic, J.; Mao, Y.; Shao, Y.; Ponder, J.; Head-Gordon, T.; Head-Gordon, M.; Skylaris, C.-K. TINKTEP: A Fully Self-Consistent, Mutually Polarizable QM/MM Approach Based on the AMOEBA Force Field. *The Journal of Chemical Physics* **2016**, *145*, 124106.
- (34) Nottoli, M.; Mennucci, B.; Lipparini, F. Excited State Born–Oppenheimer Molecular Dynamics through Coupling between Time Dependent DFT and AMOEBA. *Physical Chemistry Chemical Physics* **2020**, *22*, 19532–19541.
- (35) Bondanza, M.; Demoulin, B.; Lipparini, F.; Barbatti, M.; Mennucci, B. Trajectory Surface Hopping for a Polarizable Embedding QM/MM Formulation. *The Journal of Physical Chemistry A* **2022**, *126*, 6780–6789.
- (36) Song, C. State Averaged CASSCF in AMOEBA Polarizable Water Model for Simulating Nonadiabatic Molecular Dynamics with Nonequilibrium Solvation Effects. *The Journal of Chemical Physics* **2023**, *158*, 014101.
- (37) Song, C. State-Averaged CASSCF with Polarizable Continuum Model for Studying Photoreactions in Solvents: Energies, Analytical Nuclear Gradients, and Non-Adiabatic Couplings. *The Journal of Chemical Physics* **2022**, *156*, 104102.
- (38) Mao, Y.; Shao, Y.; Dziedzic, J.; Skylaris, C.-K.; Head-Gordon, T.; Head-Gordon, M. Performance of the AMOEBA Water Model in the Vicinity of QM Solutes: A Diagnosis Using Energy Decomposition Analysis. *Journal of Chemical Theory and Computation* **2017**, *13*, 1963–1979.
- (39) Hagras, M. A.; Glover, W. J. Polarizable Embedding for Excited-State Reactions: Dynamically Weighted Polarizable QM/MM. *Journal of Chemical Theory and Computation* **2018**, *14*, 2137–2144.
- (40) Curchod, B. F. E. *Quantum Chemistry and Dynamics of Excited States*; John Wiley & Sons, Ltd, 2020; Chapter 14, pp 435–467.

- (41) Ben-Nun, M.; Quenneville, J.; Martínez, T. J. Ab Initio Multiple Spawning: Photochemistry from First Principles Quantum Molecular Dynamics. *The Journal of Physical Chemistry A* **2000**, *104*, 5161–5175.
- (42) Nottoli, M.; Lipparini, F. General Formulation of Polarizable Embedding Models and of Their Coupling. *The Journal of Chemical Physics* **2020**, *153*, 224108.
- (43) Boulanger, E.; Thiel, W. Solvent Boundary Potentials for Hybrid QM/MM Computations Using Classical Drude Oscillators: A Fully Polarizable Model. *Journal of Chemical Theory and Computation* **2012**, *8*, 4527–4538.
- (44) Stålring, J.; Bernhardsson, A.; Lindh, R. Analytical Gradients of a State Average MCSCF State and a State Average Diagnostic. *Molecular Physics* **2001**, *99*, 103–114.
- (45) Field, M. J.; Bash, P. A.; Karplus, M. A Combined Quantum Mechanical and Molecular Mechanical Potential for Molecular Dynamics Simulations. *Journal of Computational Chemistry* **1990**, *11*, 700–733.
- (46) Singh, U. C.; Kollman, P. A. A Combined Ab Initio Quantum Mechanical and Molecular Mechanical Method for Carrying out Simulations on Complex Molecular Systems: Applications to the CH₃Cl + Cl⁻ Exchange Reaction and Gas Phase Protonation of Polyethers. *Journal of Computational Chemistry* **1986**, *7*, 718–730.
- (47) Das, D.; Eurenus, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. Optimization of Quantum Mechanical Molecular Mechanical Partitioning Schemes: Gaussian Delocalization of Molecular Mechanical Charges and the Double Link Atom Method. *The Journal of Chemical Physics* **2002**, *117*, 10534–10547.
- (48) Parks, J. M.; Hu, H.; Cohen, A. J.; Yang, W. A Pseudobond Parametrization for Improved Electrostatics in Quantum Mechanical/Molecular Mechanical Simulations of Enzymes. *The Journal of Chemical Physics* **2008**, *129*, 154106.

- (49) Zhang, Y.; Lee, T.-S.; Yang, W. A Pseudobond Approach to Combining Quantum Mechanical and Molecular Mechanical Methods. *The Journal of Chemical Physics* **1999**, *110*, 46–54.
- (50) Gökcan, H.; Vázquez-Montelongo, E. A.; Cisneros, G. A. LICHEM 1.1: Recent Improvements and New Capabilities. *Journal of Chemical Theory and Computation* **2019**, *15*, 3056–3065.
- (51) Loco, D.; Lagardère, L.; Cisneros, G. A.; Scalmani, G.; Frisch, M.; Lipparini, F.; Menucci, B.; Piquemal, J.-P. Towards Large Scale Hybrid QM/MM Dynamics of Complex Systems with Advanced Point Dipole Polarizable Embeddings. *Chemical Science* **2019**, *10*, 7200–7211.
- (52) Glover, W. J. Communication: Smoothing out Excited-State Dynamics: Analytical Gradients for Dynamically Weighted Complete Active Space Self-Consistent Field. *The Journal of Chemical Physics* **2014**, *141*, 171102.
- (53) Hohenstein, E. G.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. An Atomic Orbital-Based Formulation of the Complete Active Space Self-Consistent Field Method on Graphical Processing Units. *The Journal of Chemical Physics* **2015**, *142*, 224103.
- (54) McMurchie, L. E.; Davidson, E. R. One- and Two-Electron Integrals over Cartesian Gaussian Functions. *Journal of Computational Physics* **1978**, *26*, 218–231.
- (55) Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Kokkila-Schumacher, S. I. L.; Luehr, N.; Snyder, J. W., Jr.; Song, C.; Titov, A. V.; Ufimtsev, I. S.; Martínez, T. J. TeraChem: Accelerating Electronic Structure and Ab Initio Molecular Dynamics with Graphical Processing Units. *The Journal of Chemical Physics* **2020**, *152*, 224110.
- (56) Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Isborn, C. M.; Kokkila-Schumacher, S. I. L.; Li, X.; Liu, F.; Luehr, N.; Snyder, J. W.; Song, C.; Titov, A. V.;

- Ufimtsev, I. S.; Wang, L.-P.; Martínez, T. J. TeraChem: A Graphical Processing Unit-Accelerated Electronic Structure Package for Large-Scale Ab Initio Molecular Dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, e1494.
- (57) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Computational Biology* **2017**, *13*, e1005659.
- (58) Song, C.; Wang, L.-P. Modified OpenMM Branch for SA-CASSCF/AMOEBA Calculations. <https://github.com/leeping/openmm/tree/amoeba-casscf-81>.
- (59) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian~16 Revision C.01. 2016.
- (60) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *Journal of Chemical Theory and Computation* **2018**, *14*, 5273–5289.

- (61) Lu, T.; Chen, F. Multiwfn: A Multifunctional Wavefunction Analyzer. *Journal of Computational Chemistry* **2012**, *33*, 580–592.
- (62) Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *Journal of Chemical Theory and Computation* **2013**, *9*, 452–460, PMID: 26589047.
- (63) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic Improvement of a Classical Molecular Model of Water. *The Journal of Physical Chemistry B* **2013**, *117*, 9956–9972, PMID: 23750713.
- (64) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *The Journal of Physical Chemistry B* **2017**, *121*, 4023–4039.
- (65) McGill, R.; Tukey, J.; Larsen, W. Variations of Box Plots. *American Statistician* **1978**, *32*, 12–16.
- (66) England, C. G.; Ehlerding, E. B.; Cai, W. NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence. *Bioconjugate Chemistry* **2016**, *27*, 1175–1187.
- (67) Nemergut, M.; Pluskal, D.; Horackova, J.; Sustrova, T.; Tulis, J.; Barta, T.; Baatalah, R.; Gagnot, G.; Novakova, V.; Majerova, M.; Sedlackova, K.; Marques, S. M.; Toul, M.; Damborsky, J.; Prokop, Z.; Bednar, D.; Janin, Y. L.; Marek, M. Illuminating the Mechanism and Allosteric Behavior of NanoLuc Luciferase. *Nature Communications* **2023**, *14*, 7864.
- (68) Liu, F.; Liu, Y.; De Vico, L.; Lindh, R. Theoretical Study of the Chemiluminescent Decomposition of Dioxetanone. *Journal of the American Chemical Society* **2009**, *131*, 6181–6188.

- (69) Ma, Y. Elucidating the Multi-Configurational Character of the Firefly Dioxetanone Anion and Its Prototypes in the Biradical Region Using Full Valence Active Spaces. *Physical Chemistry Chemical Physics* **2020**, *22*, 4957–4966.
- (70) Wang, L.-P.; Song, C. Geometry Optimization Made Simple with Translation and Rotation Coordinates. *The Journal of Chemical Physics* **2016**, *144*, 214108.
- (71) Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *Journal of Chemical Theory and Computation* **2016**, *12*, 638–649.
- (72) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Research* **2012**, *40*, W537–W541.
- (73) Schenkmyerova, A.; Toul, M.; Pluskal, D.; Baatallah, R.; Gagnot, G.; Pinto, G. P.; Santana, V. T.; Stuchla, M.; Neugebauer, P.; Chaiyen, P.; Damborsky, J.; Bednar, D.; Janin, Y. L.; Prokop, Z.; Marek, M. Catalytic Mechanism for Renilla-type Luciferases. *Nature Catalysis* **2023**, *6*, 23–38.
- (74) Takatani, T.; Hohenstein, E. G.; Malagoli, M.; Marshall, M. S.; Sherrill, C. D. Basis set consistent revision of the S22 test set of noncovalent interaction energies. *The Journal of Chemical Physics* **2010**, *132*, 144104.
- (75) Boese, A. D. Density Functional Theory and Hydrogen Bonds: Are We There Yet? *ChemPhysChem* **2015**, *16*, 978–985.
- (76) Wu, Y.; Tepper, H. L.; Voth, G. A. Flexible Simple Point-Charge Water Model with Improved Liquid-State Properties. *The Journal of Chemical Physics* **2006**, *124*, 024503.
- (77) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and

- Testing of a General Amber Force Field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.
- (78) Pieri, E.; Lahana, D.; Chang, A. M.; Aldaz, C. R.; Thompson, K. C.; Martínez, T. J. The non-adiabatic nanoreactor: towards the automated discovery of photochemistry. *Chem. Sci.* **2021**, *12*, 7294–7307.
- (79) Liu, C.; Piquemal, J.-P.; Ren, P. AMOEBA+ Classical Potential for Modeling Molecular Interactions. *Journal of Chemical Theory and Computation* **2019**, *15*, 4122–4139.
- (80) Rackers, J. A.; Silva, R. R.; Wang, Z.; Ponder, J. W. Polarizable Water Potential Derived from a Model Electron Density. *Journal of Chemical Theory and Computation* **2021**, *17*, 7056–7084.
- (81) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. The Multi-State CASPT2 Method. *Chemical Physics Letters* **1998**, *288*, 299–306.
- (82) Shiozaki, T.; Győrffy, W.; Celani, P.; Werner, H.-J. Communication: Extended Multi-State Complete Active Space Second-Order Perturbation Theory: Energy and Nuclear Gradients. *The Journal of Chemical Physics* **2011**, *135*, 081106.
- (83) Song, C. New physical insights into the supporting subspace factorization of XMS-CASPT2 and generalization to multiple spin states via spin-free formulation. *The Journal of Chemical Physics* **2024**, *160*, 124106.
- (84) Battaglia, S.; Lindh, R. Extended Dynamically Weighted CASPT2: The Best of Two Worlds. *Journal of Chemical Theory and Computation* **2020**, *16*, 1555–1567.
- (85) Lipparini, F. General Linear Scaling Implementation of Polarizable Embedding Schemes. *Journal of Chemical Theory and Computation* **2019**, *15*, 4312–4317.