

# Mutually polarizable QM/MM model with *in situ* optimized localized basis functions

Cite as: J. Chem. Phys. **150**, 074103 (2019); <https://doi.org/10.1063/1.5080384>

Submitted: 08 November 2018 . Accepted: 28 January 2019 . Published Online: 19 February 2019

Jacek Dziedzic , Teresa Head-Gordon , Martin Head-Gordon , and Chris-Kriton Skylaris 



[View Online](#)



[Export Citation](#)



[CrossMark](#)

# Mutually polarizable QM/MM model with *in situ* optimized localized basis functions

Cite as: J. Chem. Phys. 150, 074103 (2019); doi: 10.1063/1.5080384

Submitted: 8 November 2018 • Accepted: 28 January 2019 •

Published Online: 19 February 2019



View Online



Export Citation



CrossMark

Jacek Dziedzic,<sup>1,2</sup> Teresa Head-Gordon,<sup>3,4</sup> Martin Head-Gordon,<sup>3</sup> and Chris-Kriton Skylaris<sup>1,a)</sup>

## AFFILIATIONS

<sup>1</sup>School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

<sup>2</sup>Faculty of Applied Physics and Mathematics, Gdańsk University of Technology, Gdańsk, Poland

<sup>3</sup>Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, California 94720, USA

<sup>4</sup>Department of Bioengineering, University of California, Berkeley, California 94720, USA

<sup>a)</sup>Author to whom correspondence should be addressed: [c.skylaris@soton.ac.uk](mailto:c.skylaris@soton.ac.uk)

## ABSTRACT

We extend our recently developed quantum-mechanical/molecular mechanics (QM/MM) approach [Dziedzic *et al.*, J. Chem. Phys. **145**, 124106 (2016)] to enable *in situ* optimization of the localized orbitals. The quantum subsystem is described with ONETEP linear-scaling density functional theory and the classical subsystem – with the AMOEBA polarizable force field. The two subsystems interact via multipolar electrostatics and are fully mutually polarizable. A total energy minimization scheme is employed for the Hamiltonian of the coupled QM/MM system. We demonstrate that, compared to simpler models using fixed basis sets, the additional flexibility offered by *in situ* optimized basis functions improves the accuracy of the QM/MM interface, but also poses new challenges, making the QM subsystem more prone to overpolarization and unphysical charge transfer due to increased charge penetration. We show how these issues can be efficiently solved by replacing the classical repulsive van der Waals term for QM/MM interactions with an interaction of the electronic density with a fixed, repulsive MM potential that mimics Pauli repulsion, together with a modest increase in the damping of QM/MM polarization. We validate our method, with particular attention paid to the hydrogen bond, in tests on water-ion pairs, the water dimer, first solvation shells of neutral and charged species, and solute-solvent interaction energies. As a proof of principle, we determine suitable repulsive potential parameters for water, K<sup>+</sup>, and Cl<sup>-</sup>. The mechanisms we employed to counteract the unphysical overpolarization of the QM subsystem are demonstrated to be adequate, and our approach is robust. We find that the inclusion of explicit polarization in the MM part of QM/MM improves agreement with fully QM calculations. Our model permits the use of minimal size QM regions and, remarkably, yields good energetics across the well-balanced QM/MM interface.

Published under license by AIP Publishing. <https://doi.org/10.1063/1.5080384>

## I. INTRODUCTION

Molecular dynamics (MD) is a well-established technique for simulating the structure and properties of systems at the atomic scale, with over four decades of applications in biochemistry and materials science, among other fields. The aim of MD is to predict macroscopic behavior from microscopic

interactions,<sup>1</sup> and the validity of results strongly depends on how accurately these interactions are described by the molecular mechanics (MM) potential.

The continual increase in available computational power not only extends the scope of MD to larger systems and longer time scales but also enables the refinement of MM models describing inter- and intra-molecular interactions. The last

two decades have witnessed the emergence of force fields that directly capture many-body polarization effects, setting out to circumvent well-known deficiencies of pairwise-additive, fixed point charge models.<sup>2-5</sup> Unable to directly account for polarization, fixed point charge force fields struggle to describe, e.g., the interactions of ions with  $\pi$ -electron systems or polar solutes in low-dielectric media,<sup>6</sup> and they are typically poorly transferable to environments or phases different from those that they were parameterized for, such as interfaces.<sup>7</sup>

The non-additive, many-body nature of polarization interactions makes polarizable models more involved and computationally demanding. Consequently, a variety of competing treatments of polarization exists (see Refs. 7-9 for a review): Drude oscillators,<sup>10,11</sup> fluctuating charges,<sup>12,13</sup> induced point dipoles,<sup>14-21</sup> or even induced multipoles of higher order.<sup>22</sup> The AMOEBA force field,<sup>15-18</sup> which is of particular significance to this work, describes polarization interactions using damped, induced, point dipoles, while for permanent electrostatics, it employs fixed multipoles up to a quadrupole *in lieu* of point charges.

Purely classical models, however sophisticated, cannot describe electronic properties, such as bandgaps or solvent shifts, or processes that intrinsically depend on the electronic degrees of freedom, such as bond-breaking. To properly model electronic phenomena, it becomes necessary to employ quantum-mechanical (QM) methods. In practical applications, density functional theory (DFT) is arguably the most commonly used approach,<sup>23</sup> owing to its relatively low computational cost. Even so, length scales (~10-100 nm) and timescales (~1  $\mu$ s) typically used in classical MD simulations remain beyond the scope of DFT today.

QM/MM combines the quantum and classical descriptions, exploiting the fact that the properties of interest are often localized to a part of the system that can be described quantum mechanically, such as a molecule, embedded in an environment that can be described more approximately with MM, e.g., a solvent. Since the seminal work of Warshel and Levitt,<sup>24</sup> a profusion of QM/MM approaches has been proposed, targeting different types of systems and varying in the level of sophistication (see, e.g., Refs. 25-36). Even a brief review of QM/MM methods is beyond the scope of this paper; however, we refer interested readers to reviews of QM/MM methods and applications in enzymology,<sup>37</sup> biochemistry,<sup>38</sup> and materials science.<sup>39</sup>

Recent improvements in force fields promptly become integrated into QM/MM methodologies, and several approaches combining QM with polarizable forcefields (dubbed QM/MMpol) have already been proposed. Many of those approaches employ induced dipoles<sup>40-55</sup> to model polarization, and others use fluctuating charges<sup>56-59</sup> or Drude oscillators.<sup>60,61</sup> Typically, the full QM density is used for all electrostatic QM/MM interactions,<sup>5,52,55</sup> but using auxiliary multipolar representations of QM (for efficiency or convenience) has also been proposed.<sup>51,62</sup> Several groups have developed models specifically focused on electronic excitations, using polarizable embedding alongside time-dependent

density functional theory (TDDFT),<sup>42,43,45,46,50,54,57,59,63</sup> where dynamic mutual polarization poses an additional challenge.<sup>46</sup>

We recently presented<sup>51</sup> a novel QM/MM approach (TINKTEP), which combines the DFT methodology of ONETEP<sup>64,65</sup> and the polarizable force field AMOEBA,<sup>15,17,18,66</sup> as implemented in TINKER.<sup>16</sup> In the TINKTEP approach, the QM and MM subsystems are coupled electrostatically and undergo mutual polarization. The electrostatic effect of the MM subsystem is included in the QM Hamiltonian, polarizing the QM subsystem by deforming its electronic charge density. Conversely, the electric field of the QM subsystem is included in the direct field that drives the polarization of the MM subsystem. A total energy minimization scheme is employed for the Hamiltonian of the coupled QM/MM system. A distinguishing feature of our approach is the use of linear-scaling DFT<sup>65,67</sup> to describe the QM subsystem with the aim of, ultimately, undertaking QM/MMpol calculations with QM regions spanning thousands of atoms.

The main limitation of our first TINKTEP model, as presented in Ref. 51, was its use of fixed localized orbitals, which represented a trade-off between simplicity and energy accuracy. In this work, we describe an extension of TINKTEP to the case where the localized orbitals are optimized *in situ*. The rationale for using optimized orbitals is the near-complete-basis-set accuracy that they offer, even when a minimal basis is used. The resultant accuracy is comparable or superior to even very large bases with fixed orbitals.<sup>68</sup>

Incorporating *in situ* orbital optimization requires computing gradients of all energy terms with respect to the expansion coefficients of the localized orbitals. We present the relevant derivation and describe how the calculation can be implemented to run in linear-scaling time. We subsequently focus on the difficulties that arise as a consequence of using optimized orbitals - QM overpolarization and unphysical charge transfer from QM to MM - and discuss workable solutions to these two problems, using simple QM/MM systems to illustrate our points. We finish by demonstrating the stability, robustness, and accuracy of our model on a number of test cases. We arrive at a robust, mutually polarizable QM/MM model with linear-scaling QM cost, which we show to be more accurate than a non-polarizable QM/GAFF (Generalized AMBER Force Field) approach, not only in terms of reducing the electrostatic disruption to the QM subsystem but also in terms of improved energetics across the QM/MM interface.

This paper is organized as follows. In Sec. II, we recount the original (fixed-orbital) TINKTEP approach. In Sec. III, we outline the generalization to *in situ* optimized orbitals and describe the additional steps that we found to be necessary for obtaining a well-behaved method. The additional steps are best justified using case studies, which, in the interest of clarity of discussion, we relegated to Appendixes A and B. Section IV is devoted to validation and demonstration of the utility of the proposed approach on a number of carefully selected systems. Conclusions and closing remarks are found in Sec. V.

## II. METHOD

## A. Conventions and notation

We follow the sign convention where electrons are positively charged. Atomic units are used throughout the text, unless indicated otherwise. Quantities typeset in bold denote Cartesian column vectors (positions  $\mathbf{r}$ , electric fields  $\mathbf{E}$ , dipoles  $\boldsymbol{\mu}$ , etc.) or Cartesian tensors of rank 2 (e.g.,  $\mathcal{T}^{\text{d-d}}_{LM}$ ). Matrices with dimensions other than  $3 \times 3$  are typeset with blackboard capitals (e.g.,  $\mathbb{K}$ ). Indices A, B, and C always refer to atoms in the QM subsystem, and indices L and M refer to atoms in the MM subsystem. Localized orbitals are indexed with Greek symbols. By van der Waals interactions, we will mean the sum of the repulsive and dispersive terms, referring to the attractive term simply as “dispersion.”

## B. Initial (fixed-orbital) TINKTEP approach

We begin the exposition of the method by briefly recounting the general idea behind TINKTEP – our first QM/MM approach proposed in Ref. 51. The system is separated into a QM subsystem and an MM subsystem, with the assumption that the separation does not cut through covalent bonds. The total energy of the coupled system is given by

$$\mathcal{E} = \mathcal{E}^{\text{QM}} + \mathcal{E}^{\text{MM}} + \mathcal{E}^{\text{QM/MM}}. \quad (1)$$

The QM region is described by the density matrix formulation of DFT in the pseudopotential approximation,

$$\begin{aligned} \mathcal{E}^{\text{QM}} = & \int \left[ -\frac{1}{2} \nabla_{\mathbf{r}}^2 \rho(\mathbf{r}, \mathbf{r}') \right]_{\mathbf{r}'=\mathbf{r}} d\mathbf{r} + \int v_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} \\ & + \frac{1}{2} \iint \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + \mathcal{E}_{\text{XC}}^{\text{QM}}[n] \\ & + \frac{1}{2} \sum_A \sum_{B \neq A}^{N_{\text{QM}}} \frac{Z_A Z_B}{|\mathbf{R}_B - \mathbf{R}_A|} + \mathcal{E}_{\text{disp}}^{\text{QM}}(\{\mathbf{R}_A\}), \end{aligned} \quad (2)$$

with the above terms describing, respectively, the kinetic energy of valence electrons, the Coulombic energy of valence electrons in the (pseudo)potential  $v_{\text{ext}}(\mathbf{r})$  of the ionic cores, the Hartree energy, the exchange–correlation energy, the mutual Coulombic interaction of  $N_{\text{QM}}$  cores having charges  $\{Z_I\}$  and positions  $\{\mathbf{R}_I\}$ , and empirical dispersion–correction. Open boundary conditions have been assumed.  $\rho(\mathbf{r}, \mathbf{r}')$  is the density matrix given by

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \varphi_{\alpha}(\mathbf{r}) K^{\alpha\beta} \varphi_{\beta}^*(\mathbf{r}'), \quad (3)$$

where  $\varphi_{\alpha}(\mathbf{r})$  are non-orthogonal generalized Wannier functions (NGWFs),<sup>69</sup> which are strictly localized within atom-centered spherical regions.  $\mathbb{K} = [K^{\alpha\beta}]$ , termed the density kernel, is the matrix representation of the density matrix in the duals of the NGWFs. The electronic (pseudo)density  $n(\mathbf{r})$  is given by

$$n(\mathbf{r}) = \rho(\mathbf{r}, \mathbf{r}), \quad (4)$$

where we assumed a closed-shell system in the interest of brevity. The last term,  $\mathcal{E}_{\text{disp}}^{\text{QM}}$ , is an empirical dispersion–correction term, which accounts for the well-known deficiency of generalized gradient approximation (GGA) DFT in describing dispersion interactions.<sup>70</sup> The exact expression depends on the model used, but the general form is that of a double sum of pairwise terms. This work uses the Elstner<sup>71</sup> formulation, with parameters determined by Hill and Skylaris.<sup>70</sup>

The MM subsystem is described by the AMOEBA<sup>18</sup> polarizable force–field, as implemented in the TINKER<sup>16</sup> code, with the following general energy expression:

$$\mathcal{E}^{\text{MM}} = \mathcal{E}_{\text{perm}}^{\text{MM}} + \mathcal{E}_{\text{pol}}^{\text{MM}} + \mathcal{E}_{\text{val}}^{\text{MM}} + \mathcal{E}_{\text{vdW}}^{\text{MM}}, \quad (5)$$

with the four energy components accounting for permanent electrostatic interactions, polarization, short-range valence interactions, and van der Waals interactions, respectively.

$\mathcal{E}_{\text{perm}}^{\text{MM}}$  is a sum of purely Coulombic multipolar interactions between atoms in the MM subsystem, with scaling factors<sup>17</sup> used to attenuate or eliminate interactions between first-, second-, third-, and fourth-nearest neighbors (as determined by bond connectivity). The full expression is given in Ref. 18, Eqs. (1) and (10). The  $\mathcal{E}_{\text{perm}}^{\text{MM}}$  term is fully local to the MM subsystem that is to say it is insensitive to the presence of the coupling with a QM subsystem.

$\mathcal{E}_{\text{pol}}^{\text{MM}}$  is the polarization energy of the MM subsystem given by [cf. Ref. 51, Eq. (A1)]

$$\begin{aligned} \mathcal{E}_{\text{pol}}^{\text{MM}} = & \frac{1}{2} \sum_L^{N_{\text{MM}}} \sum_M^{N_{\text{MM}}} \boldsymbol{\mu}_L^T \mathbf{T}_{LM} \boldsymbol{\mu}_M - \sum_L^{N_{\text{MM}}} \mathbf{E}_L^T \boldsymbol{\mu}_L \\ = & -\frac{1}{2} \sum_L^{N_{\text{MM}}} \boldsymbol{\mu}_L^T \mathbf{E}_L, \end{aligned} \quad (6)$$

where  $\mathbf{E}_L$  is the direct electric field at site L,  $\boldsymbol{\mu}_L$  is the dipole induced at site L in response to the total electric field, and  $\mathbf{T}_{LM}$  is a  $3 \times 3$  coupling tensor between sites L and M,

$$\mathbf{T}_{LM} = \begin{cases} -\mathcal{T}_{LM}^{\text{d-d}}, & L \neq M \\ \alpha_L^{-1} \mathbf{I}, & L = M \end{cases}. \quad (7)$$

Here,  $\mathcal{T}_{LM}^{\text{d-d}}$  is the Thole-damped, Cartesian dipole–dipole interaction tensor between induced dipoles at sites L and M [cf. Ref. 51, Eq. (28)]. Thole damping<sup>72–74</sup> is a modification to Coulombic electrostatics that helps prevent mutual positive feedback loops involving induced point dipoles, known as a polarization catastrophe.

In a purely MM calculation, the direct electric field  $\mathbf{E}_L$  is simply the electric field due to the permanent multipoles of MM sites. In a mutually polarizable QM/MM calculation, the direct electric field includes an additional contribution arising from the multipoles representing the QM subsystem. We give the full expression in Ref. 51, Eqs. (26)–(31), where we also explain in detail how a classical multipolar representation of a distributed QM density is obtained (Sec. II D therein). We stress that the non-additive nature of polarization means

that the polarization of MM cannot be separated into additive terms due to QM and MM, and so the entire polarization of the MM subsystem is included in  $\mathcal{E}_{\text{pol}}^{\text{MM}}$ , which explains the absence of an  $\mathcal{E}_{\text{pol}}^{\text{QM/MM}}$  term.

$\mathcal{E}_{\text{val}}^{\text{MM}}$  denotes all short-range valence interactions local to the MM subsystem. The detailed expressions for these terms can be found in Ref. 18, Eqs. (2)–(6), and we shall refrain from recounting them here.

$\mathcal{E}_{\text{vdW}}^{\text{MM}}$  accounts for van der Waals (dispersion-repulsion) interactions local to the MM subsystem. AMOEBA uses the Halgren formulation<sup>75</sup> of the buffered 14-7 potential,

$$\mathcal{E}_{\text{vdW}}^{\text{MM}}(R_{ij}) = \varepsilon_{ij} \left( \frac{1 + \delta}{\rho_{ij} + \delta} \right)^7 \left( \frac{1 + \gamma}{\rho_{ij}^7 + \gamma} - 2 \right), \quad (8)$$

where  $\rho_{ij} = R_{ij}/R^0$ ,  $\delta = 0.07$ , and  $\gamma = 0.12$ . The parameters of the potential are  $R^0$  and  $\varepsilon$ . Mixing rules for obtaining pairwise values of the parameters and a description of nuances surrounding hydrogen atoms (“reduction factors”) can be found in Ref. 18, Eqs. (7) and (8).

The final term in (1),  $\mathcal{E}^{\text{QM/MM}}$ , accounts for all interactions between the two subsystems, *except for mutual polarization*. As pointed out earlier, QM contributions to MM polarization have already been included in  $\mathcal{E}_{\text{pol}}^{\text{MM}}$  because they are not separable from intra-MM polarization. The effect of MM polarizing QM is automatically included in  $\mathcal{E}^{\text{QM}}[n(\mathbf{r})]$  by the deformation of the electronic density  $n(\mathbf{r})$  in response to the electric field of the MM subsystem.

The coupling is described by

$$\mathcal{E}^{\text{QM/MM}} = \mathcal{E}_{\text{perm}}^{\text{QM/MM}} + \mathcal{E}_{\text{vdW}}^{\text{QM/MM}}, \quad (9)$$

where the first term accounts for the electrostatic coupling between QM and permanent MM multipoles and the second term accounts for dispersion-repulsion interactions between QM and MM. In our model, the electrostatic coupling involves the full QM charge density interacting with the Coulombic (not damped) potential of the permanent MM charges, dipoles and quadrupoles,

$$\mathcal{E}_{\text{perm}}^{\text{QM/MM}} = \int v_{\text{p}}^{\text{MM}}(\mathbf{r}) n^{\text{QM}}(\mathbf{r}) d\mathbf{r}, \quad (10)$$

where

$$n^{\text{QM}}(\mathbf{r}) = n(\mathbf{r}) + \sum_A^{N_{\text{QM}}} \delta(\mathbf{r} - \mathbf{R}_A) Z_A, \quad (11)$$

and the expression for  $v_{\text{p}}^{\text{MM}}(\mathbf{r})$  can be found elsewhere [Ref. 51, Eqs. (37) and (38)].

For the QM/MM van der Waals interaction,  $\mathcal{E}_{\text{vdW}}^{\text{QM/MM}}$ , our original model uses the same classical, pairwise model that is used for MM/MM [cf. (8)], except the repulsive wall is softened slightly by using  $\delta = 0.21$  (cf. Ref. 51, Sec. III.B.3). This has the advantage of being straightforward but has two disadvantages. First, it requires choosing vdW parameter values

for atoms in the QM subsystem. Second, and more importantly, this classical form is insensitive to the electronic density of the QM subsystem and as such its contributions to electronic density gradients vanish. This means it fails to provide the Pauli repulsion that would otherwise prevent electrons from unphysically collapsing onto MM atoms. We address this issue in the revised model presented in this paper.

### III. THEORY

The main limitation of the original model described briefly above, and in detail in Ref. 51, was that the localized orbitals  $\{\varphi_\alpha\}$  were kept fixed and only the density kernel  $K^{\alpha\beta}$  was optimized. Allowing  $\{\varphi_\alpha\}$  to be optimized *in situ* constitutes the main improvement in our revised model.

#### A. *In situ* optimized NGWFs

Optimizing the NGWFs necessitates deriving and implementing functional derivatives of energy terms with respect to the NGWFs. Out of all the energy terms that make up the total energy (1), the following terms are specific to our QM/MM model and have no counterparts in standard ONETEP:  $\mathcal{E}_{\text{val}}^{\text{MM}}$ ,  $\mathcal{E}_{\text{vdW}}^{\text{MM}}$ ,  $\mathcal{E}_{\text{perm}}^{\text{MM}}$ ,  $\mathcal{E}_{\text{vdW}}^{\text{QM/MM}}$ ,  $\mathcal{E}_{\text{perm}}^{\text{QM/MM}}$ , and  $\mathcal{E}_{\text{pol}}^{\text{MM}}$ . The first three of these are local to the MM subsystem and do not depend on the electronic degrees of freedom, their derivatives with respect to the NGWFs thus vanish.  $\mathcal{E}_{\text{vdW}}^{\text{QM/MM}}$  is, so far, described by an electron-independent classical pairwise sum, so its derivative similarly vanishes. We postpone the generalization of this term to an electron-dependent form until later in the text. The remaining terms are  $\mathcal{E}_{\text{perm}}^{\text{QM/MM}}$  and  $\mathcal{E}_{\text{pol}}^{\text{MM}}$ , which we will consider now in sequence.

$\mathcal{E}_{\text{perm}}^{\text{QM/MM}}$  is given by [cf. (10) and (11)]

$$\mathcal{E}_{\text{perm}}^{\text{QM/MM}} = \int v_{\text{p}}^{\text{MM}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + \int v_{\text{p}}^{\text{MM}}(\mathbf{r}) \sum_A^{N_{\text{QM}}} \delta(\mathbf{r} - \mathbf{R}_A) Z_A d\mathbf{r}. \quad (12)$$

The potential of permanent MM multipoles  $v_{\text{p}}^{\text{MM}}$  does not depend on the electronic degrees of freedom, so the second term above does not contribute to the derivative. The first term represents an interaction of an electronic density with a fixed potential and so has the same form as the second term in (2). Thus it can be accounted for using usual ONETEP algorithms by simply adding  $v_{\text{p}}^{\text{MM}}(\mathbf{r})$  to  $v_{\text{ext}}(\mathbf{r})$ .

The term due to  $\mathcal{E}_{\text{pol}}^{\text{MM}}$  is more complicated since it involves differentiating the transformation from the QM density to the set of multipoles representing the QM subsystem that takes part in QM/MM polarization interactions. We refer the reader to Ref. 51, Sec. I.D.2 for a detailed exposition of this transformation, recounting here only the basics needed in the derivation.

All pairs (products) of overlapping NGWFs  $\varphi_\alpha$  and  $\varphi_\beta$  are expanded in terms of truncated spherical waves  $f_s(\mathbf{r})$  centered on both NGWFs, with the coefficients of the expansion given by

$$\begin{aligned}
 C_{AB}^s &= \sum_t \left( \sum_{\alpha \in A} \sum_{\beta \in B} \varphi_\alpha(\mathbf{r}) K^{\alpha\beta} \varphi_\beta^*(\mathbf{r}) \right) \Big|_{f_t} V_{AB}^{ts} \\
 &= \sum_{\alpha \in A} \sum_{\beta \in B} K^{\alpha\beta} \sum_t \left( \varphi_\alpha(\mathbf{r}) \varphi_\beta^*(\mathbf{r}) \right) \Big|_{f_t} V_{AB}^{ts} \\
 &= \sum_{\alpha \in A} \sum_{\beta \in B} K^{\alpha\beta} c_{\alpha\beta}^s, \tag{13}
 \end{aligned}$$

where  $s$  and  $t$  index the spherical waves originating on both atoms (of which there are  $N_f$  in total),  $V_{AB}^{ts}$  is an element of the inverse electrostatic overlap matrix between spherical waves originating on atoms  $A$  and  $B$ , and the notation  $\alpha \in A$  used in the summations is taken to mean “all NGWFs  $\alpha$  belonging to atom  $A$ .”

An alternative way to index the spherical waves, and in turn the coefficients  $C_{AB}^s$  and  $c_{\alpha\beta}^s$ , is via their angular, magnetic, and radial numbers:  $l$ ,  $m$ , and  $q$  and a selector for the site on which the spherical wave originates (1 for the first atom from the subscript or 2 for the second atom). This indexing scheme is useful when using the expansion to calculate the spherical multipole moments  $\{\mathcal{M}_{lm}(A)\}$  that constitute the classical representation of the density,

$$\mathcal{M}_{lm}(A) = \sum_{\substack{B \\ S_{AB} \neq 0}} \sum_q C_{AB}^{lmq,1} J_{lq}, \tag{14}$$

where  $C_{AB}^{lmq,1}$  correspond to  $C_{AB}^s$  originating only on atom  $A$  and  $J_{lq}$  is a radial, analytical integral given in Ref. 51, Eq. (21). The notation  $\sum_{\substack{B \\ S_{AB} \neq 0}}$  is taken to mean “atoms  $B$  whose NGWFs overlap with those of atom  $A$ .”

The interaction energy between all multipoles representing the QM density and the induced MM dipoles is given by

$$\mathcal{E}_{\text{pol}}^{\text{MM}} = \text{tr}[\mathbb{K}\mathbb{P}], \tag{15}$$

where  $\mathbb{P}$  is the polarization matrix, with matrix elements  $\alpha \in A, \beta \in B$  given by

$$P_{\alpha\beta} = -\frac{1}{2} \sum_{lq} J_{lq} \sum_m \left( c_{\alpha\beta}^{lmq,1} w_A^{lm} + c_{\alpha\beta}^{lmq,2} w_B^{lm} \right) = \frac{\partial \mathcal{E}_{\text{pol}}^{\text{MM}}}{\partial K^{\beta\alpha}}, \tag{16}$$

where  $w_A^{lm}$  captures the electrostatic effect of the entire system of MM induced dipoles on the QM site at  $\mathbf{R}_A$ :  $w_A^{00}$  is the Thole-damped electrostatic potential of MM induced dipoles interacting with the charge at  $\mathbf{R}_A$ ,  $\{w_A^{1,-1}, w_A^{1,0}, w_A^{1,1}\}$  is their Thole-damped electric field interacting with the dipole at  $\mathbf{R}_A$ , and  $\{w_A^{1,-2}, \dots, w_A^{1,2}\}$  is their Thole-damped electric field derivative (in spherical representation) interacting with the quadrupole at  $\mathbf{R}_A$ .

Taking the functional derivative of (15), we obtain [cf. Ref. 76, Eq. (7.25)]

$$\frac{\delta \mathcal{E}_{\text{pol}}^{\text{MM}}}{\delta \varphi_\gamma^*(\mathbf{r})} = 2 \frac{\delta P_{\alpha\beta}}{\delta \varphi_\gamma^*(\mathbf{r})} K^{\beta\alpha} + 2 \frac{\delta K^{\beta\alpha}}{\delta \varphi_\gamma^*(\mathbf{r})} P_{\alpha\beta}, \tag{17}$$

where the Einstein convention has been used for repeated Greek indices.

The non-vanishing derivative  $\frac{\delta K^{\beta\alpha}}{\delta \varphi_\gamma^*(\mathbf{r})}$  appearing in the second term results from the use of a so-called purifying transformation in ONETEP. It has already been derived [cf. Ref. 76, Eq. (7.29) or Ref. 77, Eq. (4.4.6)] and implemented in ONETEP and as such can be omitted from further discussion. Instead, we focus on the more interesting first term, involving the quantity  $\frac{\delta P_{\alpha\beta}}{\delta \varphi_\gamma^*(\mathbf{r})}$ . From (16), we obtain

$$\frac{\delta P_{\alpha\beta}}{\delta \varphi_\gamma^*(\mathbf{r})} = -\frac{1}{2} \sum_{lq} J_{lq} \sum_m \frac{\delta}{\delta \varphi_\gamma^*(\mathbf{r})} \left( c_{\alpha\beta}^{lmq,1} w_A^{lm} + c_{\alpha\beta}^{lmq,2} w_B^{lm} \right), \tag{18}$$

where we have used the fact that  $J_{lq}$  is independent of the NGWFs. The quantities  $w_A^{lm}$  and  $w_B^{lm}$  involve induced MM dipoles and, as these induce in response to a combined QM + MM electric field, they depend on the NGWFs. However, a zero residual condition at induced dipole self-consistency

$$\forall_L \frac{d\mathcal{E}_{\text{pol}}^{\text{MM}}}{d\boldsymbol{\mu}_L} = \mathbf{0} \tag{19}$$

obviates the need to calculate  $\frac{\delta \boldsymbol{\mu}_i}{\delta \varphi_\gamma^*(\mathbf{r})}$  [compare Ref. 51, Eq. (33)], allowing us to only consider the dependence of  $c_{\alpha\beta}^{lmq,1}$  and  $c_{\alpha\beta}^{lmq,2}$  on the NGWFs, and so

$$\frac{\delta P_{\alpha\beta}}{\delta \varphi_\gamma^*(\mathbf{r})} = -\frac{1}{2} \sum_{lq} J_{lq} \sum_m \left[ \frac{\delta c_{\alpha\beta}^{lmq,1}}{\delta \varphi_\gamma^*(\mathbf{r})} w_A^{lm} + \frac{\delta c_{\alpha\beta}^{lmq,2}}{\delta \varphi_\gamma^*(\mathbf{r})} w_B^{lm} \right]. \tag{20}$$

The remaining functional derivative  $\frac{\delta c_{\alpha\beta}^{lmq,1}}{\delta \varphi_\gamma^*(\mathbf{r})}$  can be calculated as follows:

$$\begin{aligned}
 \frac{\delta c_{\alpha\beta}^{lmq,1}}{\delta \varphi_\gamma^*(\mathbf{r})} &= \sum_t \frac{\delta}{\delta \varphi_\gamma^*(\mathbf{r})} \left( \varphi_\alpha(\mathbf{r}) \varphi_\beta^*(\mathbf{r}) \Big|_{f_t} \right) V_{AB}^{t,lmq,1} \\
 &= \delta_\beta^\gamma \varphi_\alpha(\mathbf{r}) \sum_t v_t(\mathbf{r}) V_{AB}^{t,lmq,1}, \tag{21}
 \end{aligned}$$

where  $v_t(\mathbf{r})$  is the potential of a spherical wave (for which an analytical expression is available).  $\frac{\delta c_{\alpha\beta}^{lmq,2}}{\delta \varphi_\gamma^*(\mathbf{r})}$  is calculated analogously.

By combining (17), (20), and (21) and expressing the sums in the first term explicitly, we obtain

$$\begin{aligned}
 \frac{\delta \mathcal{E}_{\text{pol}}^{\text{MM}}}{\delta \varphi_\gamma^*(\mathbf{r})} &= - \sum_{\substack{A \\ S_{AC} \neq 0}} \sum_{\alpha \in A} K^{\gamma\alpha} \sum_{lq} J_{lq} \sum_m \varphi_\alpha(\mathbf{r}) \\
 &\times \left[ w_A^{lm} \sum_t v_t(\mathbf{r}) V_{AC}^{t,lmq,1} + w_C^{lm} \sum_t v_t(\mathbf{r}) V_{AC}^{t,lmq,2} \right] \\
 &+ 2 \frac{\delta K^{\beta\alpha}}{\delta \varphi_\gamma^*(\mathbf{r})} P_{\alpha\beta}, \tag{22}
 \end{aligned}$$

where atom  $C$  is the host to the NGWF with respect we differentiate, i.e.,  $\gamma \in C$ . The presence of both  $w_C^{lm}$  and  $w_A^{lm}$  in (22) indicates that the gradient with respect to a particular NGWF  $\gamma \in C$  depends not only on the electrostatic effect of MM's

induced dipoles at  $\mathbf{R}_C$  but also at all centres of overlapping NGWFs  $\mathbf{R}_A$ . This is a consequence of the two-center spherical wave expansion scheme used in ONETEP.

To maintain linear scaling, an implementation must be able to evaluate (22) in  $\mathcal{O}(1)$  time since this calculation must be repeated for all NGWFs  $\gamma$ , and the number of NGWFs is proportional to  $N_{QM}$ . Our implementation in ONETEP does this by re-ordering (22) as

$$\frac{\delta \mathcal{E}_{\text{pol}}^{\text{MM}}}{\delta \varphi_{\gamma}^*(\mathbf{r})} = - \sum_{\substack{A \\ S_{AC} \neq 0}}^{N_{QM}} U_{AC}(\mathbf{r}) \sum_{\alpha \in A} K^{\gamma\alpha} \varphi_{\alpha}(\mathbf{r}) + 2 \frac{\delta K^{\beta\alpha}}{\delta \varphi_{\gamma}^*(\mathbf{r})} P_{\alpha\beta}, \quad (23)$$

where

$$U_{AC}(\mathbf{r}) = \sum_{lq} J_{lq} \sum_m \left[ w_A^{lm} \sum_t^{N_f} v_t(\mathbf{r}) V_{AC}^{t,lmq,1} + w_C^{lm} \sum_t^{N_f} v_t(\mathbf{r}) V_{AC}^{t,lmq,2} \right] \quad (24)$$

only needs to be evaluated in the intersection of the localization spheres of A and C [cf. (22)]. The cost of evaluating  $U_{AC}$  for a single pair of atoms A-C is system-size independent ( $\mathcal{O}(1)$ ) and only depends on the quality of the SW basis set. For any particular atom C, the number of atoms A whose localization spheres overlap with it plateaus at a constant that depends on the density of the system, even if the system size  $N_{QM}$  is increased to arbitrarily large values. This is made explicit by the A-C overlap condition in the first summation of (22). This means that each evaluation of (22) has  $\mathcal{O}(1)$  cost and, with  $\mathcal{O}(N)$  such operations, the approach is linear-scaling.

## B. Increased polarization damping

Polarizable force-field models that rely on the induced point dipole approximation have to contend with what is known as a *polarization catastrophe*. This well-known<sup>78</sup> artifact consists in an unbounded mutual polarization of two nearby sites through positive feedback and reflects the breakdown of the point-dipole model at a short range. The polarization catastrophe is typically mitigated by replacing Coulombic interactions involving induced dipoles with interactions that are suitably attenuated at a short range using schemes such as Thole damping.<sup>72-74</sup> This is the case in AMOEBA<sup>15</sup> and in our model.<sup>51</sup> The intensity of the damping depends on the polarizabilities of the two atoms – i.e., interactions involving atoms that polarize more readily are more aggressively damped. Beyond several Å, the difference between the Thole-damped and Coulombic quantities (potential, electric field, electric field derivative) becomes negligible and the correct long-range behavior is recovered.

The rationale for using optimized NGWFs in our model, and in ONETEP in general, is the near-complete-basis-set accuracy that they offer, comparable or superior to even very large bases with fixed orbitals.<sup>68</sup> However, this additional flexibility results in the basis becoming more diffuse, which is problematic in the context of distributed multipole analysis (DMA<sup>79,80</sup>) that we employ to obtain the multipole representation QM\* [cf. (13)–(16)]. More diffuse bases are known<sup>81,82</sup> to engender instabilities in the DMA procedure and lead to increased

charge penetration errors (CPE) due to discrepancies between the potential of the original density and that of the multipolar expansion although improved approaches have recently been proposed.<sup>82,83</sup> Indeed, our initial tests revealed that once the NGWFs are no longer fixed and are allowed to change shape during the SCF process, our QM/MM model becomes prone to a QM/MM analog of polarization catastrophe, whereby the QM subsystem becomes excessively polarized by a nearby MM site and vice versa. The problem is particularly severe for MM sites carrying a charge (ions) as they provide a larger initial polarization of QM. We devote a section in the [Appendix A](#) to an elucidation of this mode of failure using a  $\text{H}_2\text{O}:\text{Cl}^-$  system as an example. In the same section, we show that a simple increase in the damping of QM/MM polarization interactions is sufficient to prevent the QM/MM polarization catastrophe.

## C. Repulsive MM potential

Like most QM/MM models, our initial model used a classical, atom-pairwise description of QM/MM dispersion-repulsion (vdW) interactions [cf. (8)]. This strictly classical description has the disadvantage of being insensitive to the electronic degrees of freedom in the QM subsystem that is to say the QM/MM vdW energy only depends on the positions and species of the atoms. The most striking manifestation of this deficiency is that electrons in the QM subsystem do not experience any Pauli repulsion from MM sites. This can be especially problematic when the MM site is a cation, whose electrostatic potential attracts the QM electrons. With no Pauli repulsion to balance this attraction, unphysical charge transfer from QM to MM takes place.

A number of approaches have been proposed to circumvent the problem (see, e.g., Refs. 84–87), but not in the context of linear-scaling QM methods, where it becomes particularly problematic. This is because the spilled electrons accumulate near the peripheries of the localization regions, disrupting SCF convergence, which assumes localized orbitals to be well-decayed at the truncation point.

We refer the reader to the [Appendix B](#) for a case study of this undesired effect on a  $\text{H}_2\text{O}:\text{K}^+$  system, where we also demonstrate the feasibility and accuracy of an improved model which eliminates this issue. The improvement consists in replacing the repulsive term of Halgren's vdW potential with a density overlap model<sup>86</sup> that is sensitive to the QM electronic degrees of freedom, accounting for QM/MM Pauli repulsion. We retain Halgren's classical description for QM/MM dispersion interactions.

In an overlap model, the Pauli repulsion energy is assumed to be proportional to the overlap between densities, i.e.,

$$\mathcal{E}_{\text{rep}}^{\text{QM/MM}} = \int n(\mathbf{r}) \sum_L^{N_{MM}} \kappa_L n_L^{\text{MM}}(\mathbf{r} - \mathbf{R}_L) d\mathbf{r}, \quad (25)$$

where  $n(\mathbf{r})$  is the QM electronic density,  $n_L^{\text{MM}}(\mathbf{r})$  is a model density centered on MM atom L, and  $\kappa_L$  is a proportionality constant with a suitable unit.

A reasonable model density is that of a 1s Slater-type function

$$n_L^{\text{MM}}(\mathbf{r}) = |\psi_{1s,L}(\mathbf{r})|^2 = \frac{\zeta_L^3}{\pi} e^{-2\zeta_L r}. \quad (26)$$

Instead of working with model densities, we can think of MM atoms as equipped with a model repulsive electrostatic potential, leading to an equivalent energy expression

$$\epsilon_{\text{rep}}^{\text{QM/MM}} = \int n(\mathbf{r}) \sum_L^{N_{\text{MM}}} v_L^{\text{MMrep}}(\mathbf{r} - \mathbf{R}_L) d\mathbf{r} \quad (27)$$

together with an equivalent MM repulsive potential

$$v_L^{\text{MMrep}}(\mathbf{r}) = \frac{\kappa_L \zeta_L^3}{\pi} e^{-2\zeta_L r} = \frac{A_L}{\pi} e^{-2\zeta_L r}, \quad (28)$$

characterized by two parameters – a magnitude  $A$  and an inverse-width  $\zeta$ , both of which depend on the chemical species of MM atom  $L$ .

The form of (27) is that of a static external potential acting on the electronic density, which means  $v^{\text{MMrep}}(\mathbf{r}) = \sum_L^{N_{\text{MM}}} v_L^{\text{MMrep}}(\mathbf{r} - \mathbf{R}_L)$  can simply be added to  $v_{\text{ext}}(\mathbf{r})$  in (2), and no new energy gradient expressions need to be derived for this term. Unless stated otherwise, all results presented in this paper have been obtained with the model that includes the MM repulsive potential (and excludes the repulsive contribution from the Halgren QM/MM vdW expression).

Regarding computational efficiency, we point out that the integral in (27) only needs to be computed over the union of localization spheres of the QM subsystem (since  $n(\mathbf{r})$  vanishes elsewhere). Furthermore, and more importantly, since  $v_L^{\text{MMrep}}(\mathbf{r})$  decays exponentially, only those regions of the QM subsystem that are within a short cutoff radius (say, 5 Å) from any MM atom need to be considered. Generating  $v_L^{\text{MMrep}}(\mathbf{r})$  in a sphere around  $\mathbf{R}_L$ , with the sphere radius system-size independent, is an  $\mathcal{O}(1)$  operation for a single MM atom  $L$ . The number of MM atoms within a cutoff radius from the QM subsystem will be proportional to the surface area of the QM subsystem and so to  $N_{\text{QM}}^{2/3}$ . The total cost of evaluating (27) thus scales  $\mathcal{O}(N_{\text{QM}}^{2/3})$ .

Naturally, physically reasonable values for  $A_L$  and  $\zeta_L$  need to be determined for all species of interest appearing in the MM subsystem. As a proof of concept, in Sec. IV A, we show how suitable values can be found for  $\text{Cl}^-$ ,  $\text{K}^+$ , and  $\text{H}_2\text{O}$ .

## IV. RESULTS

In this section, we demonstrate the accuracy and viability of the proposed approach on a number of systems. In all QM calculations, we used the Perdew–Burke–Ernzerhof (PBE)<sup>88</sup> exchange–correlation functional, with an empirical dispersion correction in the Elstner<sup>71</sup> formulation, with parameters determined by Hill and Skylaris.<sup>70</sup> The NGWF localization radius was set to 3.7 Å.

### A. Interaction energy curves

We begin by examining the interaction energy curves of three simple systems:  $\text{H}_2\text{O}:\text{K}^+$ ,  $\text{H}_2\text{O}:\text{Cl}^-$ , and a water dimer.

The latter two systems were studied in our earlier studies,<sup>5,52</sup> using a different QM/MM model and using energy decomposition analysis (EDA) to compare AMOEBA against a high-quality DFT functional  $\omega\text{B97X-V}$ .<sup>89</sup> For each of the systems, we compare the predictions of the QM/MM model that is the focus of this paper and those of AMOEBA, against reference results obtained from fully QM calculations (i.e., PBE-D as described above). All QM calculations used a kinetic energy cutoff of 1290 eV.

By performing a parameter scan in the space of  $\{A, \zeta\}$ , we established MM repulsive potential parameters for  $\text{K}^+$  that, for this system, are optimal in the sense of minimizing the mean squared difference between the interaction energy curves from QM/MM and fully QM calculations. The values we obtained are  $A_{\text{K}^+} = 230 \text{ Ha/e}$  and  $\zeta_{\text{K}^+} = 1.379 a_0^{-1}$ . The interaction energy curves are compared in Fig. 1. AMOEBA (green curve) is seen to model this interaction faithfully, with the position of the minimum accurate to 0.007 Å and only very slight underbinding (less than 1 kcal/mol). This degree of agreement is expected since the charge density of  $\text{K}^+$  is tightly localized and thus well-approximated by a point multipole model, with very little charge penetration error. The QM  $\text{H}_2\text{O}/\text{MM K}^+$  description is in even better agreement with full QM – the position of the minimum is accurate to 0.002 Å and the energy is no further than 0.3 kcal/mol from the fully QM result for all interatomic separations.

We now turn our attention to the  $\text{H}_2\text{O}:\text{Cl}^-$  system, which we expect to be more difficult for a polarizable point dipole model due to the larger electronic delocalization of  $\text{Cl}^-$ , which increases the charge penetration error. By following the same protocol as for the  $\text{H}_2\text{O}:\text{K}^+$  system, we established optimal parameters for the repulsive MM potential for  $\text{Cl}^-$ :  $A_{\text{Cl}^-} = 250 \text{ Ha/e}$  and  $\zeta_{\text{Cl}^-} = 1.140 a_0^{-1}$ , which, compared to  $\text{K}^+$ , represent a marginally stronger and somewhat less localized potential, consistent with expectations.

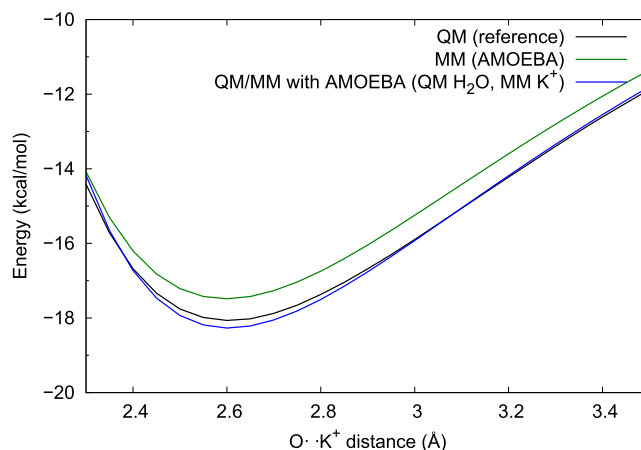


FIG. 1. Interaction energy for  $\text{H}_2\text{O}:\text{K}^+$  system as a function of the  $\text{O} \cdots \text{K}^+$  distance. QM/MM results (blue) are compared against full QM (black). Pure MM results obtained with AMOEBA are shown in green.



The interaction energy curves are compared in Fig. 2. Compared to our QM reference, AMOEBA is seen to underbind at all interatomic separations, particularly at short distances, where the magnitude of the error increases from  $\approx 1$  kcal/mol to over 5 kcal/mol. A large fraction of this error can be attributed to the neglect of charge transfer. In our reference, fully QM calculations as much as  $0.22e$  is transferred from the  $\text{Cl}^-$  ion to the water molecule at the shortest studied separation ( $2.8 \text{ \AA}$ ), corresponding to a stabilizing effect of  $\approx -5$  kcal/mol. As the separation is increased, this charge transfer becomes less pronounced – at  $4 \text{ \AA}$  only  $0.05e$  is transferred and the corresponding change in energy is only about  $-0.2$  kcal/mol.

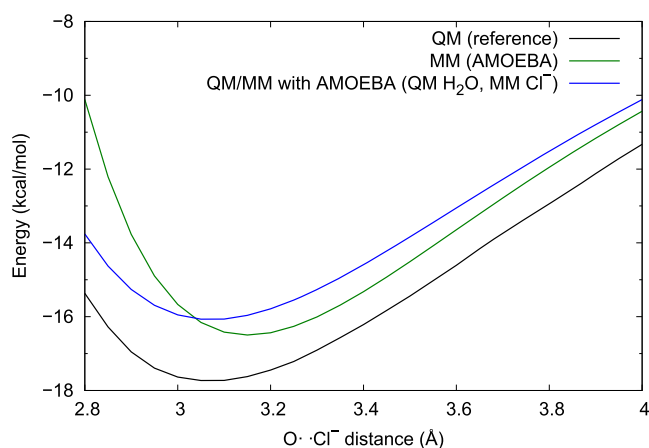
However, we must acknowledge the fact that the QM reference curve against which AMOEBA is benchmarked is too a result of a physical model. DFT calculations involve a number of approximations, chief among which are the use of an approximate exchange–correlation functional, the pseudopotential approximation, the use of a finite basis, and – in linear-scaling DFT – the use of finite radii for the localized orbitals. Different choices for these parameters will lead to slightly, but noticeably, different interaction energy curves, particularly since the water molecule is well-known to be difficult to describe with GGA DFT (see, e.g., Refs. 90–92). For instance, when instead of PBE,  $\omega\text{B97X-V}^{89}$  is used for the same system (as reported by some of us in Ref. 5, cf. Fig. 3 therein), the reference curve shifts upwards by  $\approx 1$  kcal/mol in the long range (practically matching AMOEBA), and by as much as  $\approx 3.2$  kcal/mol at  $2.8 \text{ \AA}$ , reducing AMOEBA's perceived underbinding at the shortest separation studied here to 2 kcal/mol. We thus caution against treating all differences between MM and QM reported here strictly as deficiencies of the MM model.

Naturally, the charge penetration error is also expected to be more significant for  $\text{Cl}^-$  than for  $\text{K}^+$ . AMOEBA does not explicitly model charge transfer or account for charge

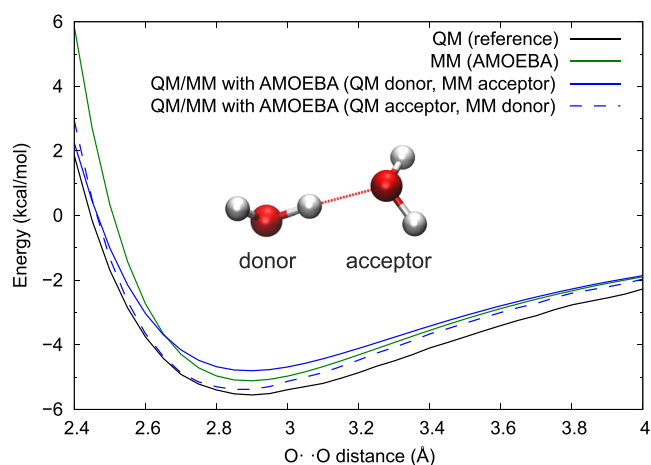
penetration and must resort to approximating these effects through polarization and vdW interactions. More severe underbinding at short distances leads to a shift in the position of the minimum, which, compared to our QM reference, AMOEBA overestimates by  $0.08 \text{ \AA}$  (or  $0.05 \text{ \AA}$  against the QM reference of Ref. 5). Our QM/MM model achieves better agreement with fully QM results, underbinding by less than 2 kcal/mol, with the magnitude of the error being almost independent of the distance between  $\text{Cl}^-$  and the water molecule. Thus, the predicted interaction energy curve is very similar in shape to the reference one, only shifted by a constant, and the position of the minimum is predicted very accurately (to  $0.002 \text{ \AA}$ ), showing the QM/MM interface to be well-balanced in this scenario.

We now turn our attention to the  $\text{H}_2\text{O}$  dimer. In earlier work<sup>52</sup> on the same system, we showed that charge penetration is significant at the equilibrium distance and below it, making this system challenging for AMOEBA, which has to compensate for CPE by artificially softening the repulsive vdW wall, relying on cancellation of errors to model the hydrogen bond. Thus (cf. Fig. 3), the agreement between AMOEBA and a fully QM calculation worsens at short separations, where AMOEBA underbinds by as much as 4 kcal/mol (2.9 kcal/mol against the QM reference of Ref. 5), but is still remarkably good at the equilibrium distance and beyond, where AMOEBA underbinds by only  $\approx 0.5$  kcal/mol. The rms error across the entire curve is 1.1 kcal/mol. The position of the minimum is also predicted accurately (to  $0.004 \text{ \AA}$ ).

Determining suitable parameters for our model's repulsive MM potential for O and H atoms is more challenging than in the previous two cases. First, there are four parameters to be simultaneously optimized ( $A_{\text{H}}$ ,  $\zeta_{\text{H}}$ ,  $A_{\text{O}}$ ,  $\zeta_{\text{O}}$ ), making the parameter scan more involved. Second, if our model is to be well-transferable, it must accurately describe both the situation where the hydrogen bond donor is described by QM (and the acceptor by MM), and the situation where QM is



**FIG. 2.** Interaction energy for a  $\text{H}_2\text{O}-\text{Cl}^-$  system as a function of the  $\text{O}\cdots\text{Cl}^-$  distance. QM/MM results (blue) are compared against full QM (black). Pure MM results obtained with AMOEBA are shown in green.



**FIG. 3.** Interaction energy for a water dimer as a function of the  $\text{O}\cdots\text{O}$  distance. QM/MM results (blue) are compared against full QM (black). Pure MM results obtained with AMOEBA are shown in green.

used to describe the hydrogen bond acceptor (and MM – the donor).

After a thorough parameter scan, we determined the following suitable values for the parameters of the MM repulsive potential:  $A_H = 35 \text{ Ha}/e$ ,  $\zeta_H = 2.40 a_0^{-1}$ ,  $A_O = 550 \text{ Ha}/e$ ,  $\zeta_O = 1.58 a_0^{-1}$ . With these values, the interaction energy curves predicted by our QM/MM model (Fig. 3, blue curves) are in very good agreement with fully QM results. When QM is used to model the hydrogen bond donor (solid blue curve), the rms error in energy across the entire curve is 0.7 kcal/mol and the position of the minimum is accurate to 0.002 Å. When QM is used to model the hydrogen bond acceptor (dashed blue curve), the rms error in energy is only 0.4 kcal/mol, but the position of the minimum is predicted less accurately and is underestimated by 0.026 Å. Crucially, in both cases, the interaction energy curve stays within 1 kcal/mol from the reference curve obtained with fully QM calculations, even at separations below the equilibrium distance. This indicates that the QM/MM interface in our model is well-balanced even in the presence of hydrogen bonds.

The above examination of the performance of our QM/MM model for three representative systems (MM cation, MM anion, and MM neutral molecule with a hydrogen bond spanning the QM/MM interface) can be considered a proof of concept. We showed that our QM/MM model is stable for all studied intermolecular separations, even well below the equilibrium distance, and that it gives reasonable predictions for interaction energy profiles, which we find remarkable given that in the studied systems, the crucial interactions crossed the QM/MM boundary.

In all examples so far, we used the H<sub>2</sub>O molecule for the QM subsystem and so the question of whether our QM/MM model is transferable, particularly concerning the parameterization of the MM repulsive potential, remains open. In the text that follows, we will examine the model's performance for a number of different molecules, both neutral and charged, demonstrating that it is indeed transferable as its predictions remain accurate.

The need to determine suitable parameters for the repulsive MM potential of all MM species of interest can be seen as a weakness of our model. Intuitively, one would hope that the parameters  $A_L$  and  $\zeta_L$  could be derived from corresponding classical vdW parameters  $\varepsilon_L$  and  $R_L^0$  – e.g., we expected  $A_L \sim \varepsilon_L$  and  $\zeta_L \sim 1/R_L^0$ . However, we found this not to be the case. For instance, for the parameters, we determined  $\frac{A_H}{A_O} \approx 15.7$ , whereas  $\frac{\varepsilon_H}{\varepsilon_O} \approx 8.1$ , that is, our model uses a substantially weaker potential on H. Similarly, we have  $\frac{\zeta_H}{\zeta_O} \approx 1.52$ , whereas  $\frac{1/R_H^0}{1/R_O^0} \approx 1.28$ , meaning the potential on H used in our model is also slightly tighter.

One reason is that while  $v_L^{\text{MMrep}}(\mathbf{r})$  [cf. (28)] is linear in  $A_L$  just like  $\varepsilon_{\text{vdW}}^{\text{MM}}(R_{ij})$  [cf. (8)] is linear in  $\varepsilon_L$ , the energy expression (27) for  $\varepsilon_{\text{rep}}^{\text{QM/MM}}$  is not linear in  $A_L$ . This is because  $n(\mathbf{r})$  implicitly depends on  $A_L$ , that is to say, the electronic density responds to the MM repulsive potential by deforming

accordingly. Thus, not only is  $\varepsilon_{\text{rep}}^{\text{QM/MM}}$  not linear in  $A_L$ , but also other energy terms are indirectly influenced by the repulsive MM potential through the change in  $n(\mathbf{r})$ . Another reason is that the vdW parameters adopted in AMOEBA have been fitted to partially compensate for the deficiencies in the classical treatment of electrostatics, some of which are no longer present in our QM/MM formulation. Finally, AMOEBA employs additional “tweaks” in its vdW formulation, for instance, the repulsive sites of H atoms are slightly offset from the actual atomic sites (“reduction factor”).

While we plan to investigate routes for automatically obtaining  $A_L$  and  $\zeta_L$  in future work, in this paper, we will focus on MM water, for which we have obtained good parameters already. In this way, we can apply our model to a large class of systems that is of practical interest – QM solutes embedded in MM water. We defer applications with different MM species to a later time.

## B. Interaction energies of solutes with water shells of increasing size

We now set out to demonstrate the transferability of our model, turning our attention to a number of QM solutes embedded in spherical shells of MM water. We will use the same systems and the same methodology as in our earlier work<sup>51</sup> – the QM subsystem will only encompass the solute, and we will study the behavior of the QM/MM system as the size of the MM H<sub>2</sub>O shell is increased (cf. Fig. 3 in Ref. 51).

Three of the solutes were chosen from the SAMPL4 blind challenge:<sup>93</sup> (a) (–)-menthol, (b) diphenylhydramine, and (c) 2-chloro-4-hydroxy-3,5-dimethoxybenzaldehyde. These moderately sized molecules (31, 40, and 23 atoms, respectively) encompass a number of chemical features: a cyclohexane ring (a), an ether group (b), an aromatic ring (b), an amine group (b), a halogen atom (c), and an aldehyde group (c). The remaining three molecules were (d) ammonia (NH<sub>3</sub>), (e) the ammonium ion (NH<sub>4</sub><sup>+</sup>), and (f) the cyanide ion (CN<sup>–</sup>) – were chosen with the aim of verifying if our model correctly describes small and charged solutes.

We compared four computational approaches:

- Fully QM calculations with no embedding (entire system treated at the DFT level of theory), which serve as reference;
- QM calculations using a purely electrostatic embedding, where the QM subsystem encompassed only the solute, and H<sub>2</sub>O molecules were described with fixed partial charges. In this setup, only a fixed, external potential is included in the QM Hamiltonian; we emphasize the neglect of vdW interactions between the QM and the embedding;
- QM/MM calculations with either a fixed point-charge embedding (GAFF v1.5<sup>94</sup>) or a polarizable embedding (AMOEBA). Here too the QM subsystem encompassed only the solute, and all water molecules were described by a classical force field. For the fixed point-charge (GAFF) embedding, vdW interactions between the solvent and solute were included at the MM level of

theory (Lennard-Jones potential). Thus, the MM repulsive potential introduced in Sec. III C was not used in this case. Similarly, polarization damping (Sec. III B) was not relevant here as the force field was not polarizable. For the polarizable embedding (AMOEBE), we used the final, refined QM/MM model, as described in Secs. III B–III C;

- (d) Fully MM calculations, where the entire system was treated classically (with GAFF or AMOEBA).

In fixed point charge QM/MM calculations and in QM calculations with fixed point-charge embedding, we used partial charges of  $0.417e$  for H atoms and  $-0.834e$  for O atoms, which are identical to the TIP3P<sup>95</sup> model used in GAFF. All QM calculations used a kinetic energy cutoff of 1000 eV. The configurations were prepared by solvating the solutes in approximately 660 explicit H<sub>2</sub>O molecules under periodic boundary conditions. Classical polarizable MD trajectories in the  $NpT$  ensemble ( $p = 1$  atm,  $T = 298$  K) were then obtained, and the final configurations after 50 ps were used. Final densities were  $1.00 \pm 0.01$  g cm<sup>-3</sup>. For details on how the configurations were processed for calculations under open boundary conditions, details of cutoffs assumed, and the models and parameters used for the solutes and water, see Ref. 51, Sec. 3.B.1.

We begin by comparing the interaction energies between the QM solute and the MM water shell as a function of the size of the shell (number of H<sub>2</sub>O molecules). To better elucidate the long-range behavior, in Fig. 4, we only plot the error in the energy with respect to the fully QM calculation that we use as reference. Even though the systems studied here are the same as in our earlier work,<sup>51</sup> we point out that the energy error curves are not directly comparable with those of Ref. 51, since in current work, we used a more refined QM approach with *in situ* optimized NGWFs both for QM/MM calculations and for the fully QM reference.

Our first observation is that for all six systems, purely MM calculations with AMOEBA (green circles) clearly outperform GAFF (gray diamonds). With the exception of NH<sub>3</sub> and CN<sup>-</sup>, typical errors in the GAFF description are at least three times larger than their AMOEBA counterparts. Even in cases where GAFF fares relatively well (NH<sub>3</sub>), or where AMOEBA's error is rather large (CN<sup>-</sup>), long-range behavior is clearly much better described by AMOEBA. This can be appreciated from the much flatter profiles of the AMOEBA curves, which indicate an almost constant energy shift from the reference QM calculation. The energy changes from adding subsequent water molecules are more erratic for GAFF, and the convergence with the number of H<sub>2</sub>O molecules is much worse. As expected, this is particularly pronounced for charged solutes – in the case of CN<sup>-</sup>, for instance, the binding energy is not well converged even at 400 MM H<sub>2</sub>O molecules. We attribute this to polarization partly compensating for the boundary effects that result from truncating the water shells.

The behavior of point-charge electrostatic embedding (QM + EE, red diamonds) is best understood by comparing it against QM/MM with GAFF (orange crosses) since these

two approaches only differ by the neglect of QM/MM vdW interactions in the former. This neglect leads to a very rapid accumulation of error at a short range, particularly for larger solutes (a)–(c), where this interaction is more significant. At longer range, QM/MM vdW interactions are well-decayed, which is reflected in the almost identical profiles of QM/MM-GAFF and QM + EE curves starting at approximately 150 H<sub>2</sub>O molecules for large solutes (a)–(c) and as early as approximately 50 H<sub>2</sub>O molecules for small solutes (d)–(f). As expected, the neglect of QM/MM vdW interactions makes the QM + EE approach inadequate for calculating interaction energies between the QM solute and embedding although occasionally (e.g., for NH<sub>3</sub> and CN<sup>-</sup>) the error fortuitously cancels out some of the errors in the electrostatics.

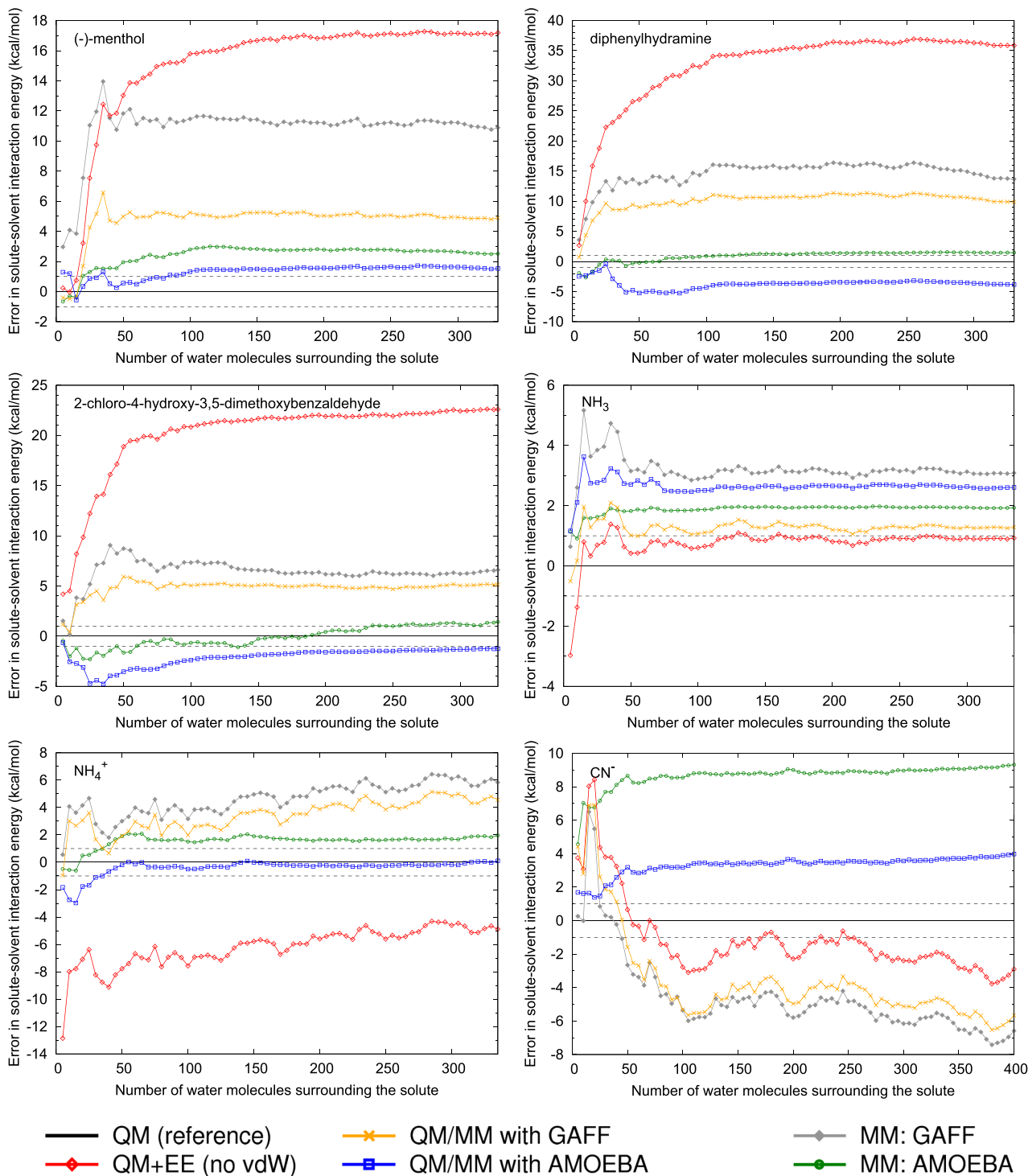
It is also worthwhile to compare the results of QM/MM-GAFF (orange crosses) against purely MM GAFF calculations (gray diamonds) because it reveals the effect of treating the solute at the QM level of theory, all other components of the two models being identical. For all six solutes, QM/MM-GAFF is more accurate, and, as the long-range profiles of the two curves are almost identical, it is clear that this gain in accuracy is due to a much improved description of short-range interactions, i.e., the ability of the QM subsystem to realistically polarize in response to the MM environment.

Of greatest interest to this paper is, of course, the comparison between QM/MM-AMOEBE (blue squares) and QM/MM-GAFF (orange crosses). In terms of absolute errors in energy, our model outperforms QM/MM-GAFF in all cases except for NH<sub>3</sub>. Moreover, the long-range behavior of QM/MM-AMOEBE is much better (flatter curves), particularly for charged solutes, where all fixed-point charge approaches (MM-GAFF, QM/MM-GAFF, QM + EE) clearly suffer from neglecting polarization. Out of all five models, the QM/MM-AMOEBE model has the lowest maximum error in the long range (4 kcal/mol for CN<sup>-</sup>, compared with 9 kcal/mol of AMOEBA, and maximum errors in excess of 10 kcal/mol for the other approaches). We summarize these results in Table I, from which it also becomes clear that, when averaged over all six systems, our approach has the lowest signed and unsigned errors of all the considered approaches.

### C. Dipole moments of solutes in water shells of increasing size

Satisfied that the energetics of our QM/MM interface is accurate, we now focus on how the QM solute is affected by the presence of the QM/MM interface. Naturally, we would like the electronic structure of the QM solute in the presence of MM embedding to resemble the electronic structure of the full QM system as much as possible, i.e., for the MM embedding to faithfully mimic QM. Since we cannot compare electronic energy levels between QM/MM and full QM and comparing electronic densities would require density partitioning, we will use the total dipole moment of the solute as a proxy.

In Fig. 5, we plot the magnitudes of the solute dipole moment for the same six solutes as a function of the size of the



**FIG. 4.** Error in the solute-solvent interaction energy with increasing number of H<sub>2</sub>O molecules surrounding the solute with reference to DFT calculation – with fixed point charge embedding (red, ◇), with GAFF embedding (orange, ×), with AMOEBA embedding (blue, □), and in purely MM calculations with GAFF (gray, ◆) and AMOEBA (green, ○). In QM/MM calculations, only the solute is included in the QM subsystem.

**TABLE I.** Comparison of interaction energy accuracy offered by fixed point charge (GAFF) and multipolar polarizable (AMOEBA) force-fields in MM calculations and QM/MM calculations and by electrostatic embedding (QM + EE). The values shown are errors (kcal/mol) with respect to DFT (PBE-D) reference calculations, averaged over systems with 200 + H<sub>2</sub>O molecules. RMSE: root mean square error, MSE: mean signed error.

Molecule	MM GAFF	MM AMOEBA	QM + EE (Point-charge)	QM/MM GAFF	QM/MM AMOEBA (this work)
(-)-Menthol	11.1	2.7	17.1	5.0	1.9
Diphenylhydramine	15.3	1.5	36.4	10.8	-3.4
2-Cl-4-OH-3,5-dimethoxy-BALD	6.2	1.0	22.2	4.9	-1.4
NH <sub>3</sub>	3.1	1.9	0.9	1.3	2.9
NH <sub>4</sub> <sup>+</sup>	5.7	1.7	-5.0	4.4	0.0
CN <sup>-</sup>	-5.8	9.0	-2.2	-4.9	4.0
RMSE	8.9	4.0	18.9	5.9	2.6
MSE	5.9	3.0	11.6	3.6	0.7

water shell. In QM/MM calculations, the solute (QM) dipoles are immediately available. In fully QM calculations, the solute dipoles were obtained from DMA analysis. In fully MM calculations, the solute dipoles are either obtained by a suitable vector summation of permanent dipoles with induced dipoles (AMOEBA) or, in the absence of polarization, are simply constant (GAFF). For neutral systems, the dipole moment is invariant to the choice of the reference point. For charged systems, and in fully QM calculations where charge transfer between the solute and solvent can make the total solute charge non-zero, we chose the centroid of the molecule as the reference point.

Our first observation is that, particularly for larger solutes, the solute dipole moment is rather sensitive to the environment and can change abruptly depending on where subsequent H<sub>2</sub>O molecules are added. The qualitative behavior of this sensitivity is captured to a similar degree by all models, except of course MM-GAFF, which does not permit solute polarization. The accuracy of the constant dipole moment of the GAFF solute is hit-and-miss – e.g., GAFF's prediction is excellent for (-)-menthol, severely underestimated for diphenylhydramine and NH<sub>4</sub><sup>+</sup>, and severely overestimated for CN<sup>-</sup>.

The predictions of QM + EE (red diamonds) and QM/MM-GAFF (orange crosses) are expected to be identical since the two approaches only differ by the absence/presence of a classical QM/MM vdW term that does not affect the electronic degrees of freedom. In practice, we observe very small differences (<0.1 D) that are the consequence of different smearing of the singularities of the Coulombic permanent fixed-point charges of the embedding on the Cartesian grid on which electronic density is evaluated in ONETEP.

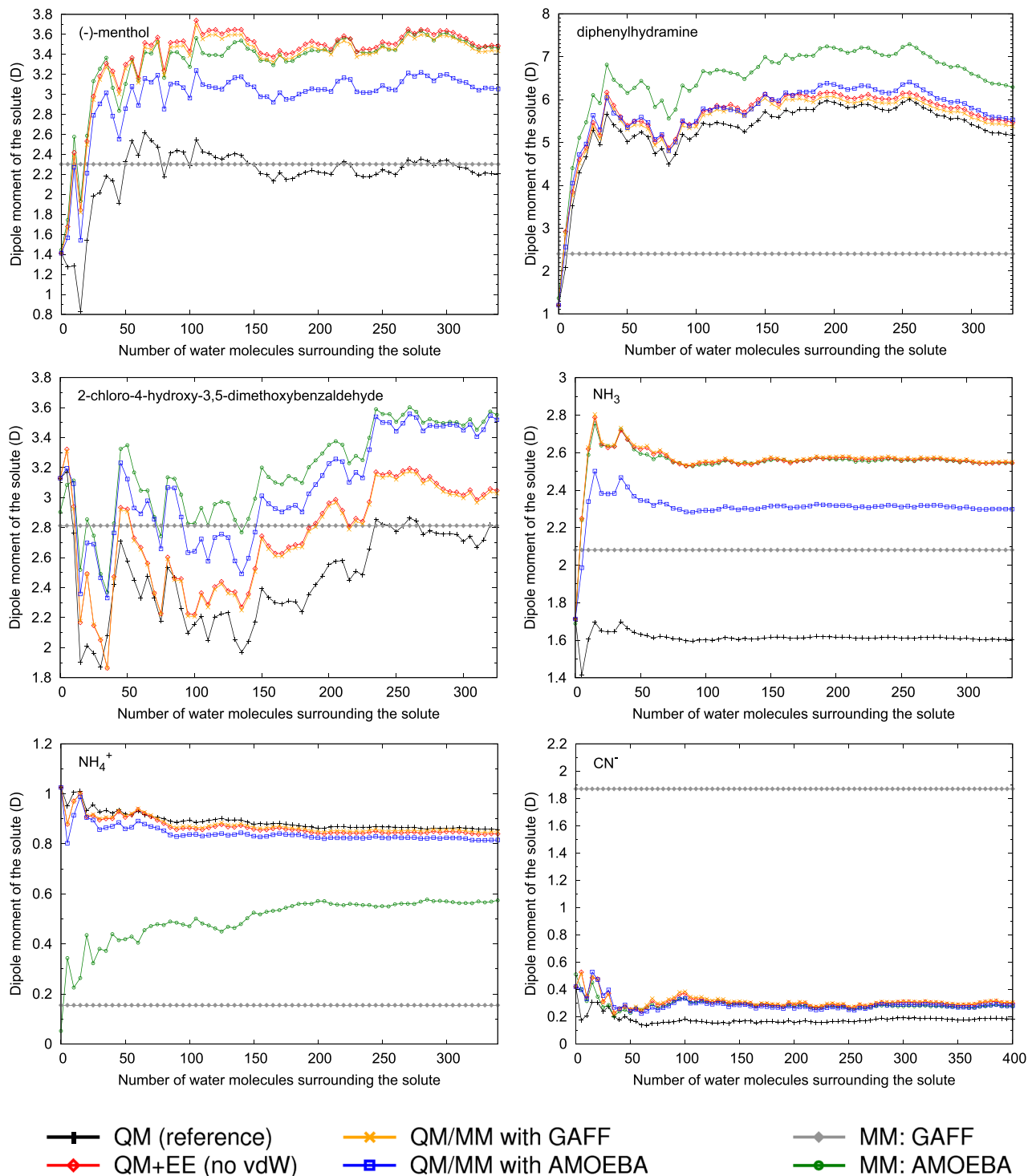
For all six solutes, the predictions of our QM/MM model are more accurate than those of AMOEBA, indicating the expected superiority of a QM description of the solute (this is most striking for NH<sub>4</sub><sup>+</sup>, which is underpolarized with AMOEBA). For three of the six solutes ((-)-menthol, NH<sub>3</sub>, CN<sup>-</sup>), the predictions of our model are closest to the fully QM results in absolute terms. For the remaining three molecules, our model

is slightly less accurate than QM/MM-GAFF, but not much so. Furthermore, this only happens when AMOEBA itself fares worse (diphenylhydramine, 2-Cl-4-OH-3,5-dimethoxy-BALD, and NH<sub>4</sub><sup>+</sup>), possibly implicating the polarizable water model, rather than the QM/MM interface, as the culprit. When the errors are averaged over all the systems, our QM/MM model yields the lowest root mean square error (RMSE). Details are summarized in Table II.

#### D. Interaction energies of solutes with 1st solvation shell

We now turn our attention to the interaction between three solutes: H<sub>2</sub>O, Cl<sup>-</sup>, Na<sup>+</sup>, and their first solvation shells. The three solutes are meant to be representative of neutral, anionic, and cationic species, respectively. To benchmark our QM/MM approach, we model only the solute at the QM level of theory, while the solvent (water) is being described by AMOEBA. We calculate solute-solvent interaction energies, comparing the performance of our approach against a non-polarizable model (QM/MM-GAFF) and purely classical models (where the entire system is described with GAFF or AMOEBA). Our aim is to verify whether our QM/MM interface correctly reproduces the purely QM results obtained at the same level of theory. Thus, our reference is pseudopotential DFT + D with PBE, the limitations of which we acknowledged earlier. We caution the reader against interpreting the discrepancies between MM results and our reference as “failures” of MM in absolute terms.

To obtain meaningful statistics, we performed calculations for 100 configurations (for each solute) obtained from MD runs, where each solute was solvated in 215 H<sub>2</sub>O molecules. In each MD snapshot, all but N<sub>1st</sub> H<sub>2</sub>O molecules closest to the solute were then stripped, leaving only the first solvation shell. The values of N<sub>1st</sub> were 4, 8, and 6 for H<sub>2</sub>O, Cl<sup>-</sup>, and Na<sup>+</sup>, respectively. The configurations studied are the same as used in our earlier work.<sup>52</sup> For a more detailed description of how the configurations were obtained and the



**FIG. 5.** Comparison of solute dipole moments with increasing number of H<sub>2</sub>O molecules surrounding the solute – with fixed point charge embedding (red, ◊), with GAFF embedding (orange, ×), with AMOEBA embedding (blue, □), and in purely MM calculations with GAFF (gray, ♦) and AMOEBA (green, ○). Reference QM calculations: black, +. In QM/MM calculations, only the solute is included in the QM subsystem.

**TABLE II.** Comparison of solute dipole moment accuracy offered by fixed point charge (GAFF) and multipolar polarizable (AMOEBA) force-fields in MM calculations and QM/MM calculations and by electrostatic embedding (QM + EE). The values shown are rms errors (D) with respect to DFT (PBE-D) reference calculations, averaged over all systems (solute + water shells). RMSE: root mean square error.

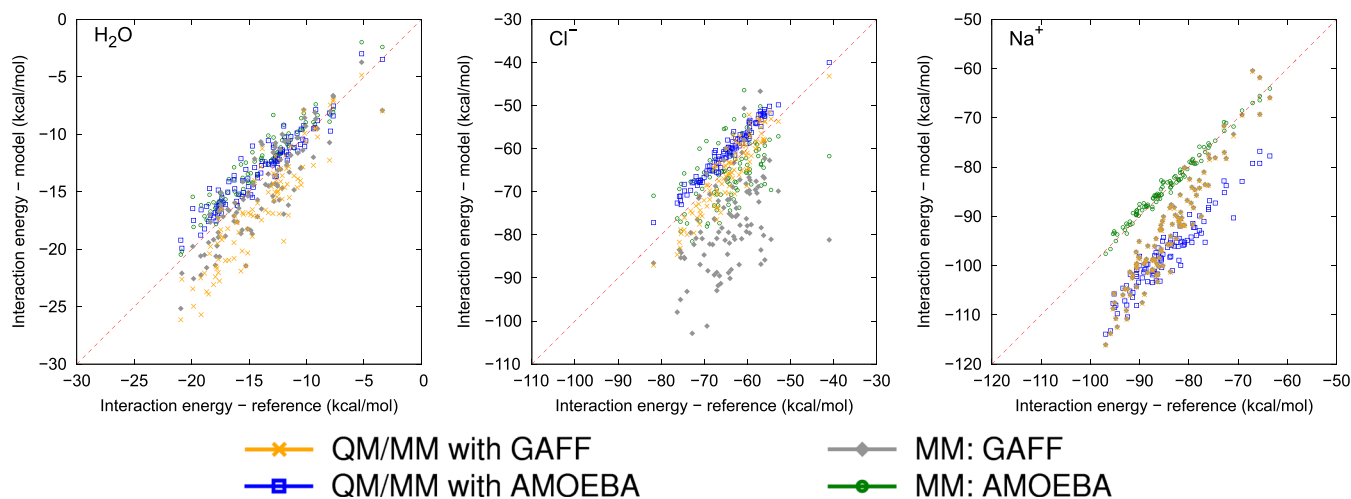
Molecule	MM GAFF	MM AMOEBA	QM + EE (Point-charge)	QM/MM GAFF	QM/MM AMOEBA (this work)
(-)-menthol	0.31	1.13	1.19	1.14	0.77
diphenylhydramine	3.00	1.18	0.32	0.24	0.38
2-Cl-4-OH-3,5-dimethoxy-BALD	0.45	0.73	0.30	0.28	0.61
NH <sub>3</sub>	0.47	0.94	0.95	0.95	0.70
NH <sub>4</sub> <sup>+</sup>	0.74	0.41	0.02	0.02	0.05
CN <sup>-</sup>	1.69	0.11	0.13	0.14	0.11
RMSE	1.47	0.84	0.65	0.63	0.52

rationale for choosing  $N_{1st}$ , we refer the reader to Ref. 52, Sec. III.C.

The calculated solute-solvent interaction energies are plotted in Fig. 6, while Table III reports crucial statistics. For the H<sub>2</sub>O solute, our QM/AMOEBA model performs well, with a general trend of underbinding by about 1 kcal/mol, and comparing favorably against QM/GAFF in all five metrics (rms error, mean signed error, maximum error, correlation coefficient, and slope of linear fit to the model's interaction energy vs. reference). We find this rather remarkable since QM/MM models are typically not very good at reproducing interaction energies spanning the interface. The largest error for our model is 3.7 kcal/mol, which is better than pure AMOEBA (4.6 kcal/mol), and much better than pure GAFF (6.2 kcal/mol) or QM/GAFF (7.3 kcal/mol). The correlation between the model and reference is also very good ( $r = 0.95$ ) although pure AMOEBA does marginally better ( $r = 0.96$ ).

The Cl<sup>-</sup> solute exposes the weaknesses of purely MM treatments. GAFF is particularly inaccurate here, with rms and mean signed errors in excess of 10 kcal/mol and a maximum error of over 40 kcal/mol, which is not surprising, given the likely importance of polarization effects in this system, for both the solute and solvent. These results were obtained using the vdW parameters proposed by Fox and Kollman:<sup>96</sup>  $R^* = 1.948 \text{ \AA}$ ,  $\epsilon = 0.265 \text{ kcal/mol}$ , which we used for consistency with previous work. We note in passing that GAFF performs slightly better here when vdW parameters tailored specifically for Cl<sup>-</sup> in TIP3P water are used (Joung and Cheatham:<sup>97</sup>  $R^* = 2.513 \text{ \AA}$ ,  $\epsilon = 0.0356 \text{ kcal/mol}$ ). The improved values are shown in parentheses in Table III.

AMOEBA's predictions are better (an rms error of 6.6 kcal/mol), but it does not avoid occasional embarrassments (a max error of 20.7 kcal/mol). Both QM/MM models perform significantly better, which highlights the importance of treating the Cl<sup>-</sup> ion at the QM level of theory, in order to



**FIG. 6.** Interaction energies (kcal/mol) between three solutes (H<sub>2</sub>O, Cl<sup>-</sup>, and Na<sup>+</sup>) and their first solvation shells – a comparison of MM and QM/MM models (y axis) with fully QM reference results (x axis) for 100 snapshots.

**TABLE III.** Comparison of accuracy offered by fixed point charge (GAFF) and multipolar polarizable (AMOEBA) force-fields in MM calculations and QM/MM calculations for the interaction energies between three solutes ( $\text{H}_2\text{O}$ ,  $\text{Cl}^-$ , and  $\text{Na}^+$ ) and their first solvation shells.  $\Delta E$  are errors in energy (kcal/mol) with respect to DFT (PBE-D) reference calculations, averaged over all 100 snapshots:  $\Delta E_{\text{rms}}$  – root mean square error,  $\Delta E_{\text{mse}}$  – mean signed error,  $\Delta E_{\text{max}}$  – maximum error.  $r$  is the Pearson correlation coefficient between  $E_{\text{model}}$  and  $E_{\text{QM}}$ ,  $a$  is the slope of linear fit of  $E_{\text{model}}$  vs.  $E_{\text{QM}}$ . Values in bold denote the most accurate model in each category, and values in italic denote the least accurate model. Values in parentheses were obtained with alternative vdW parameters for  $\text{Cl}^-$  (see text).

System		MM GAFF	MM AMOEBA	QM/MM GAFF	QM/MM AMOEBA (this work)
$\text{H}_2\text{O}-\text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	2.1	1.7	3.1	<b>1.4</b>
	$\Delta E_{\text{mse}}$	<b>-0.8</b>	1.3	-2.4	0.9
	$\Delta E_{\text{max}}$	6.2	4.6	7.3	<b>3.7</b>
	$r$	0.88	<b>0.96</b>	0.91	0.95
	$a$	<b>1.05</b>	0.91	1.17	0.85
$\text{Cl}^- - \text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	15.8 (10.7)	6.6	<b>3.9</b>	4.1
	$\Delta E_{\text{mse}}$	-12.7 (5.8)	<b>-1.1</b>	-1.4	3.9
	$\Delta E_{\text{max}}$	40.2 (31.2)	20.7	9.4	<b>7.1</b>
	$r$	0.47 (0.51)	0.53	0.92	<b>0.98</b>
	$a$	0.77 (0.82)	0.57	1.23	<b>0.97</b>
$\text{Na}^+ - \text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	10.4	<b>1.0</b>	10.5	13.5
	$\Delta E_{\text{mse}}$	-9.2	<b>0.5</b>	-9.3	-13.3
	$\Delta E_{\text{max}}$	19.2	<b>2.9</b>	19.3	19.4
	$r$	0.95	<b>0.99</b>	0.95	0.95
	$a$	1.50	0.96	1.50	<b>0.99</b>

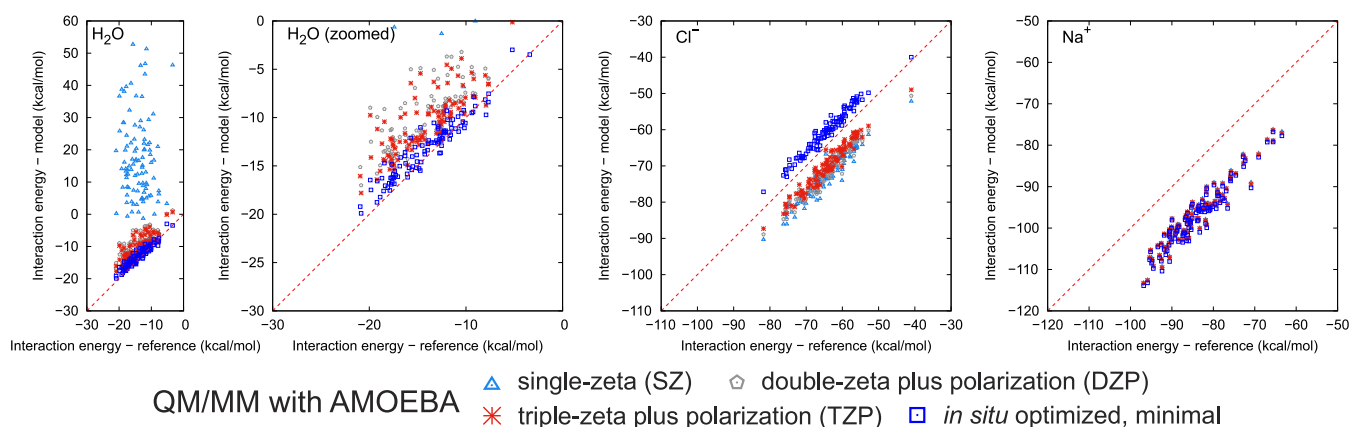
be consistent with the latter. Our QM/AMOEBA model correlates better with pure QM ( $r = 0.98$  against  $r = 0.92$  for QM/GAFF, a linear slope of 0.97 against 1.23 for QM/GAFF), but it is seen to underbind slightly across the board (rms error of 4.1 kcal/mol, compared to 3.9 kcal/mol for QM/GAFF). Its maximum error is 7.1 kcal/mol, which is rather large, but still better than QM/GAFF (9.4 kcal/mol) and much better than the double-digit errors of MM models.

For  $\text{Na}^+$  and its first solvation shell, AMOEBA performs very well, while all the remaining models are rather inaccurate. Since  $\text{Na}^+$  is a compact, barely polarizable ion, it is well-described by MM methods. This explains why GAFF results are almost identical to QM/GAFF results. What is significantly more important in this system is the description of the water solvent. GAFF's water model cannot capture the polarization of the solvent, which is highly relevant here, due to the charge on the  $\text{Na}^+$  solute. Thus GAFF and QM/GAFF both yield a poor description of the whole system, with rms errors above 10 kcal/mol and maximum errors of almost 20 kcal/mol. This is despite using vdW parameters for  $\text{Na}^+$  that were specifically tailored for a sodium ion in TIP3P water (Ref. 97,  $R^* = 1.369 \text{ \AA}$ ,  $\epsilon = 0.087439 \text{ kcal/mol}$ ). AMOEBA, by contrast, performs very well, with an rms error of only 1 kcal/mol and good correlation with purely QM results ( $r = 0.96$ ), highlighting the importance of a polarizable description of the water solvent. Since our QM/AMOEBA model shares its description of the solvent with AMOEBA, one would expect it to yield a similarly good description. However, this is not the case. While the correlation with purely QM results is good ( $r = 0.95$ , slope of 0.99), there is significant overbinding for all

snapshots, leading to large errors in energy, dominated by a mean signed error of  $-13.3 \text{ kcal/mol}$ . This almost constant shift points to a deficiency of our QM/MM interface in handling cationic solutes, presumably due to the repulsive MM potential having been parameterized only using  $\text{H}_2\text{O}-\text{H}_2\text{O}$  interactions. We attribute the observed overbinding to an insufficient repulsion between the compact  $\text{Na}^+$  and nearby MM oxygen atoms.

We will now briefly investigate the effect of using *in situ* optimized orbitals on the quality of QM/AMOEBA. We calculated the interaction energies of the three systems from Fig. 6 using fixed pseudoatomic orbitals [single-zeta (SZ), double-zeta and polarization (DZP), triple-zeta and polarization (TZP)] and compared them with results obtained using an *in situ* optimized minimal basis. We report the results in Fig. 7 and Table IV. As expected, for a compact  $\text{Na}^+$  cation, a minimal (SZ) basis is sufficient, and increasing the flexibility of the basis makes very little difference, with the slope and correlation coefficients practically unchanged, and an essentially rigid shift of the interaction energies by 0.5–0.7 kcal/mol compared to an *in situ* optimized basis. Since our model systematically overbinds this system, the resultant shift actually makes the fixed-basis results marginally better, owing to cancellation of a small fraction of the error. For a diffuse  $\text{Cl}^-$  anion, the effect of the basis size is more pronounced, with a clear systematic improvement of about 1.5 kcal/mol in the mean signed error (MSE) for each time the basis set quality is increased although the optimized basis “overshoots” by a small amount, underbinding by 3.9 kcal/mol. Nevertheless, QM/MM with an *in situ* optimised basis yields the most accurate results





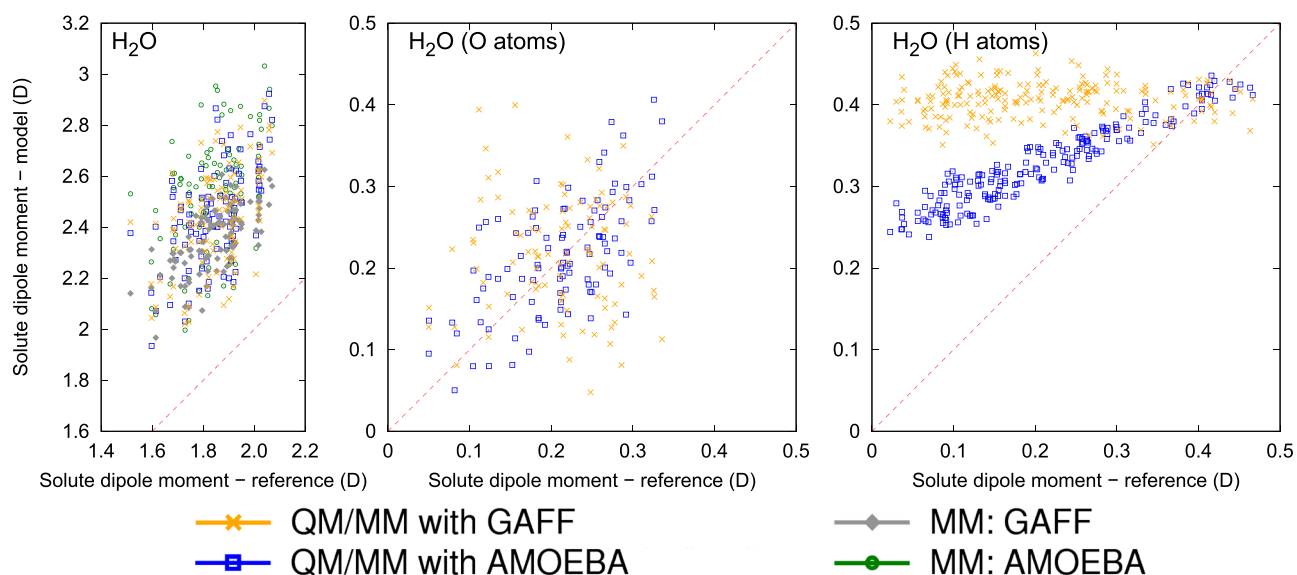
**FIG. 7.** Interaction energies (kcal/mol) between three solutes ( $\text{H}_2\text{O}$ ,  $\text{Cl}^-$ , and  $\text{Na}^+$ ) and their first solvation shells – comparison of QM/AMOEBa using fixed pseudoatomic orbital basis and *in situ* optimized minimal basis (y axis) with fully QM reference results (x axis) for 100 snapshots.

under all metrics, except for the slope (where the differences between basis set qualities are marginal). For  $\text{H}_2\text{O}$ , the effect of using an optimized basis is dramatic, which is expected, since we anticipate the orbitals in a molecule to be poorly described with small basis sets. Consequently, a minimal fixed basis (SZ) yields entirely wrong results, consistently predicting large and *positive* interaction energies, with an MSE as large as 31.2 kcal/mol. The addition of polarization functions improves results dramatically, but convergence with the size

of the basis set is slow – DZP yields an MSE of 4.2 kcal/mol, and TZP yields 3.6 kcal/mol, with correlation coefficients of only  $\sim 0.8$ . Only when *in situ* optimized orbitals are used do the results improve markedly – the MSE falls below 1 kcal/mol, the correlation coefficient exceeds 0.95, and maximum error diminishes by a factor of 2.7 compared to TZP. The optimized orbital formulation of QM/AMOEBa is a clear winner in this case, underscoring the advantages of using an *in situ* optimized basis set for the QM subsystem.

**TABLE IV.** Comparison of accuracy offered by the QM/AMOEBa model for the interaction energies between three solutes ( $\text{H}_2\text{O}$ ,  $\text{Cl}^-$ , and  $\text{Na}^+$ ) and their first solvation shells depending on the quality of the QM basis set.  $\Delta E$  are errors in energy (kcal/mol) with respect to DFT (PBE-D) reference calculations, averaged over all 100 snapshots:  $\Delta E_{\text{rms}}$  – root mean square error,  $\Delta E_{\text{mse}}$  – mean signed error,  $\Delta E_{\text{max}}$  – maximum error.  $r$  is the Pearson correlation coefficient between  $E_{\text{model}}$  and  $E_{\text{QM}}$ ,  $a$  is the slope of linear fit of  $E_{\text{model}}$  vs.  $E_{\text{QM}}$ .  $N_{\text{basis}}$  is the number of QM basis functions for the solute. Values in bold denote the most accurate model in each category, and values in italic denote the least accurate model.

System		QM/MM AMOEBa SZ	QM/MM AMOEBa DZP	QM/MM AMOEBa TZP	QM/MM AMOEBa <i>In situ</i> optimized minimal
$\text{H}_2\text{O}-\text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	34.1	4.8	4.1	<b>1.4</b>
	$\Delta E_{\text{mse}}$	31.2	4.2	3.6	<b>0.9</b>
	$\Delta E_{\text{max}}$	68.4	10.9	10.1	<b>3.7</b>
	$r$	-0.07	0.78	0.82	<b>0.95</b>
	$a$	-0.26	0.75	0.79	<b>0.85</b>
	$N_{\text{basis}}$	6	23	29	6
$\text{Cl}^- - \text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	8.6	7.4	6.1	<b>4.1</b>
	$\Delta E_{\text{mse}}$	-8.4	-7.2	-5.8	<b>3.9</b>
	$\Delta E_{\text{max}}$	13.9	12.0	10.0	<b>7.1</b>
	$r$	0.96	0.97	0.97	<b>0.98</b>
	$a$	0.99	1.00	<b>1.00</b>	0.97
	$N_{\text{basis}}$	9	13	17	9
$\text{Na}^+ - \text{H}_2\text{O}$	$\Delta E_{\text{rms}}$	<b>12.8</b>	12.9	12.9	13.5
	$\Delta E_{\text{mse}}$	<b>-12.6</b>	-12.7	-12.8	-13.3
	$\Delta E_{\text{max}}$	<b>18.1</b>	18.5	18.5	19.4
	$r$	<b>0.95</b>	0.95	0.95	0.95
	$a$	<b>0.99</b>	0.99	0.99	0.99
	$N_{\text{basis}}$	8	21	29	8



**FIG. 8.** Dipole moments (magnitudes) of the solute ( $\text{H}_2\text{O}$ ) embedded inside a first solvation shell – comparison of MM and QM/MM models (y axis) with fully QM reference results (x axis) for 100 snapshots. Left panel – total dipole moment evaluated at the centroid of the solute. Middle panel – atom-centered dipole moment of solute O atom. Right panel – atom-centered dipole moment of solute H atoms. MM GAFF and MM AMOEBA atom-centered dipoles (not shown here) cannot be meaningfully compared with QM/MM because in these MM models point charges are fixed. Furthermore, MM GAFF atom-centered dipoles are zero by construction.

### E. Dipole moments of solutes with 1st solvation shell

Having examined the energetics of our model, in the last step of our analysis, we will verify how the QM subsystem (solute) is affected by the presence of the QM/MM interface. Following the rationale of Sec. IV C, we will compare solute dipole moments as a proxy for the shape of the electronic density. As reference, we use fully QM calculations, where we partitioned the dipole moments into atomic contributions using DMA.<sup>98</sup> We compare the magnitudes of the total solute dipole between the two QM/MM approaches (where the solute dipole is simply the total QM dipole) and the two purely MM approaches [where the solute dipole is directly obtained from atomic dipoles (including induced dipoles in the case of AMOEBA) and charges].

We begin our discussion with the water pentamer system. While an isolated water molecule is charge-neutral, and its dipole moment is position-independent, in fully QM calculations of a solvated water molecule, we observe moderate charge transfer (below  $\pm 0.1e$ ) between the central  $\text{H}_2\text{O}$  molecule and the solvation shell, which makes the solute dipole position-dependent, and necessitates choosing a reference point. For consistency with the rest of this work, we chose the centroid of the solute as the point, where we evaluate the dipole moment. The total dipole moments, as well as individual atom-centered dipole moments (for O and H atoms separately) are plotted in Fig. 8 and summarized in Table V.

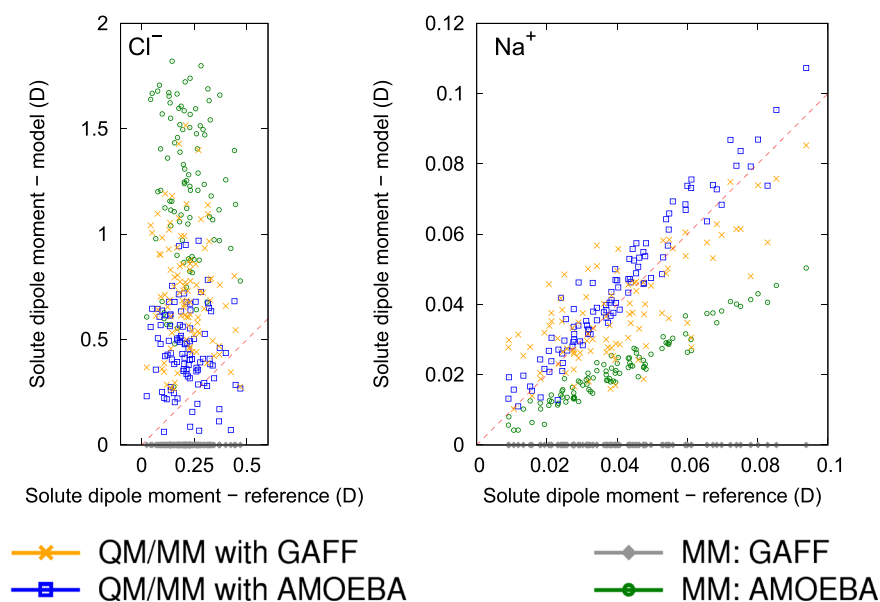
Qualitatively, the total dipole moment of the central  $\text{H}_2\text{O}$  molecule (solute) is described similarly by both MM and QM/MM approaches – with substantial scatter and

overpolarization (of about 0.6 D) relative to the purely QM results. We attribute this to the differences in the description of the four surrounding  $\text{H}_2\text{O}$  molecules – here only the reference uses a DFT description, while both MM and QM/MM approaches use a classical description. Given that standard DFT GGA models are known to struggle to correctly describe the structure and properties of water (cf., e.g., Refs. 90–92), and that we expect the dipole moment to be larger than the gas phase value (1.85 D) and lower than the value for bulk water ( $\approx 2.7$  D),<sup>99</sup> we believe that it is in fact the reference calculation that underpolarizes the solvated  $\text{H}_2\text{O}$  molecule. Neither of the MM or QM/MM models seems to have a particular advantage in this case although QM/AMOEBA correlates slightly better with the reference.

It is more interesting to examine the individual atom-centered dipoles. Here we expect MM and QM/MM results

**TABLE V.** Comparison of solute dipole moment accuracy offered by fixed point charge (GAFF) and multipolar polarizable (AMOEBA) force-fields in MM calculations and QM/MM calculations. The values shown are rms errors (debye) with respect to DFT (PBE-D) reference calculations, averaged over 100 snapshots.

Solute	MM GAFF	MM AMOEBA	QM/MM GAFF	QM/MM AMOEBA (this work)
$\text{H}_2\text{O}$	0.521	0.713	0.612	0.624
$\text{H}_2\text{O}$ (O atom)	...	...	0.099	0.059
$\text{H}_2\text{O}$ (H atoms)	...	...	0.229	0.136
$\text{Cl}^-$	0.232	1.140	0.584	0.330
$\text{Na}^+$	0.044	0.021	0.013	0.007



**FIG. 9.** Dipole moments (magnitudes) of two ionic solutes (left:  $\text{Cl}^-$ , right:  $\text{Na}^+$ ) embedded inside their first solvation shells – comparison of MM and QM/MM models (y axis) with fully QM reference results (x axis) for 100 snapshots. NB.: In MM, GAFF atom-centered dipoles are zero by construction.

to differ substantially because both MM approaches are disadvantaged by the constraint of fixed charge on the atoms, while in QM/MM, the charge density is free to transfer between atoms. GAFF, being non-polarizable, additionally yields zero atom-centered dipoles by construction. Thus, it is only meaningful here to benchmark the two QM/MM approaches against one another. For the O atom, we find QM/AMOEBA to be superior to QM/GAFF in terms of correlation with the reference result and a lower rms error (0.059 D vs 0.099 D). For the H atom, QM/GAFF does not reproduce the change in the dipole moment between the snapshots at all, while QM/AMOEBA shows the right trend although it mostly overpolarizes.

We finish with an examination of the two ionic solutes (Fig. 9, Table V). We evaluate the dipole at the position of the ion. For  $\text{Cl}^-$ , we find all approaches, except for the non-polarizable GAFF, to generally overpolarize the ion, which is again expected, given the differences between classical and DFT treatments of the water solvent. AMOEBA is the least accurate possibly because of the large magnitude of charge penetration error for a diffuse  $\text{Cl}^-$  ion. QM/AMOEBA performs better than QM/GAFF (an rms error of 0.330 D vs. 0.584 D). For  $\text{Na}^+$ , which barely polarizes at all ( $\mu < 0.1$  D), QM/AMOEBA turns out to be the most accurate (an rms error of 0.007 D vs. 0.013 D for QM/GAFF) although all models are qualitatively correct.

We conclude that, at least for small hydrated solutes, our model is superior to QM/GAFF not only in how the QM subsystem is affected by the QM/MM interface but also in better energetics across the interface.

## V. CONCLUSIONS

We presented and benchmarked a new mutually polarizable QM/MM model, where the QM subsystem is described

using DFT with *in situ* optimized, localized orbitals (non-orthogonal generalized Wannier functions, NGWFs), and the MM subsystem is described using the AMOEBA force field. By implementing our model in the ONETEP linear-scaling DFT framework, we pave the way for affordable large-scale QM/MM calculations, with QM subsystems spanning thousands of atoms. However, in this work, we only studied small QM subsystems (up to 40 atoms), which are outside the linear-scaling regime.

The rationale for optimizing NGWFs is the near-complete-basis-set accuracy that they enable even with a minimal basis. This high accuracy is comparable or superior to even very large bases with fixed orbitals.<sup>68</sup> As part of this work, we derived and implemented the necessary gradients of the total QM + MM energy with respect to the NGWFs, enabling their *in situ* optimization also in the context of QM/MM calculations. We demonstrated how the additional flexibility of an *in situ* optimized basis exacerbates known problems of polarizable QM/MM methods – catastrophic overpolarization of the QM region (particularly in the presence of MM ions) and unphysical charge transfer (“charge spilling”) from QM to MM sites (particularly in the presence of MM cations).

We developed, presented, and validated conceptually straightforward solutions to both of these issues. We demonstrated that QM overpolarization can be mitigated by modifying the value of the Thole damping parameter only for polarization interactions spanning the QM/MM interface. This suitably attenuates QM/MM polarization at a very short range while having negligible medium-range and long-range effect. We addressed the charge spilling by introducing a more refined model of QM/MM Pauli repulsion interactions. This refinement replaces classical Halgren vdW repulsion with an electrostatic repulsive potential originating on MM atoms, which is parameterized to mimic Pauli repulsion. This approach

is functionally equivalent to a density-overlap-based Pauli repulsion energy model with fixed, species-dependent densities placed on MM atoms, and, crucially, actual QM densities. This formulation is sensitive to the electronic degrees of freedom, which prevents the electronic density from excessive spilling.

Our modified approach requires parameterizing the repulsive potential with two values per MM species. We demonstrated how suitable values could be determined for  $K^+$  and  $Cl^-$  ions and for the atomic components of water, giving us confidence that this can be performed in principle. We did not identify a simple relation that would enable us to easily derive the sought parameters from classical vdW parameters, but we plan to investigate this further in future work.

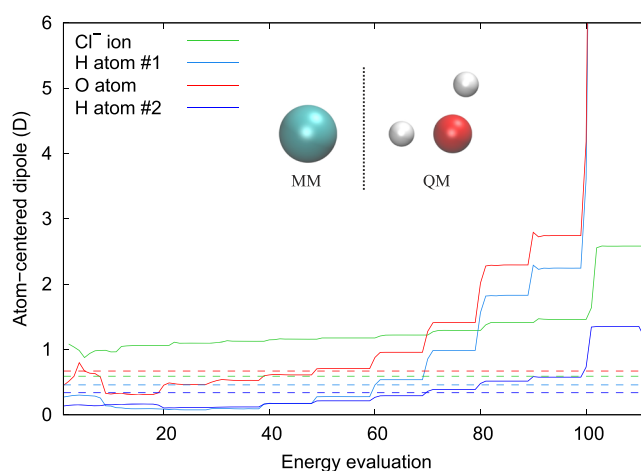
We performed extensive tests to evaluate the transferability and reliability of our model with focus on the MM treatment of water. Using a variety of molecules, from small ions and neutral systems to larger molecules (up to 40 atoms), we showed that our model is, in general, although not universally, superior to nonpolarizable QM/MM and to purely MM approaches. This was reflected in lower disruptive effect of the QM/MM interface on the QM subsystem (which we assessed by comparing the dipole moments against a fully QM reference) but also by better energetics, when calculating interaction energies between the QM and MM subsystems. We find the latter particularly promising, as QM/MM models are typically very poor at describing interaction energies across the QM/MM boundary.

## ACKNOWLEDGMENTS

The authors are grateful to Alex Albaugh and Richard Bradshaw for fruitful discussions. AMOEBA parameterizations and initial structures of solutes used in this work were kindly provided by Richard Bradshaw. Structures of water dimers and solvated water molecules and ions were kindly provided by Yuezhi Mao. J.D. and C.-K.S. acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC Grant Nos. EP/J015059/1 and EP/K039156/1) and of the UKCP consortium (EPSRC Grant No. EP/K013556/1) for access to the ARCHER national supercomputer and the IRIDIS High Performance Computing Facility of the University of Southampton. T.H.-G. and M.H.-G. acknowledge the support by Grant No. CHE-1665315 from the U.S. National Science Foundation. The majority of the calculations in this work were carried out at the TASK Academic Computer Centre in Gdansk, Poland.

## APPENDIX A: QM/MM POLARIZATION CATASTROPHE: DIAGNOSIS AND PREVENTION

In this section, we elaborate on the QM/MM polarization catastrophe (cf. Sec. III B) that can arise when NGWFs in the QM subsystem are optimized *in situ*. We use an  $H_2O:Cl^-$  system as an example. Figure 10 shows the magnitudes of atom-centered dipoles computed from the DMA procedure for QM atoms and the induced dipole of the MM  $Cl^-$  ion (whose permanent dipole is of course zero) in the course of

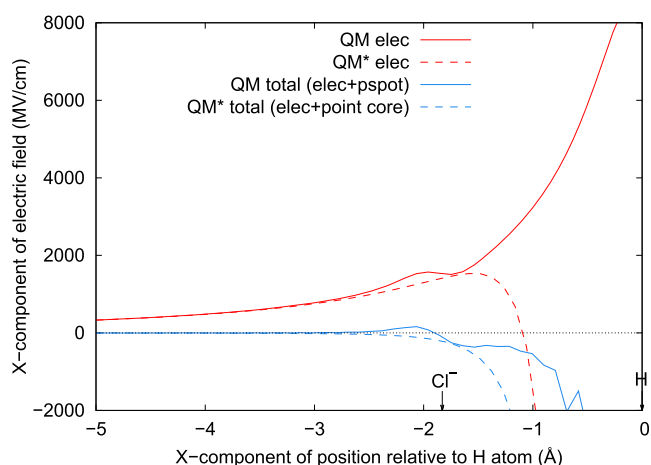


**FIG. 10.** Atom-centered dipoles (magnitudes) on atoms in the QM subsystem (red, blue) and in the MM subsystem (green) in the course of SCF optimization for a  $H_2O:Cl^-$  system (shown in the inset). In the point-dipole model under standard Thole damping, the two subsystems polarize one another to infinity. Dashed lines denote reference values obtained from a fully QM calculation.

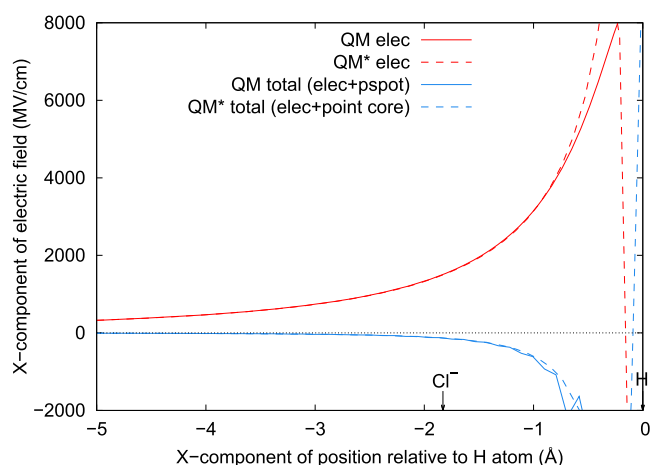
SCF optimization. A mutual positive feedback can be observed to intensify at about step 80, quickly leading to absurdly large dipole moments (in excess of 1000 D). The sharp, step-like changes to the dipole values correspond to NGWF optimization steps and the relatively flatter parts of the graph – to the density kernel optimization steps. The classical degrees of freedom (MM induced dipoles) are fully optimized for each energy evaluation. The expected dipole moments on all atoms (as calculated from a fully QM reference calculation), shown with dashed lines, are well below 1 D. It is clear that the MM site, in particular, is overpolarized from the very first step of the optimization.

The main underlying reason for this is the inaccuracy of the multipolar approximation of QM density at short range, i.e., the charge penetration error (in this system, the  $Cl^-$  ion is 1.8 Å away from the H atom, while localized NGWFs extend for 3.7 Å). This can be appreciated in Fig. 11, where we plot the relevant component of the electric field due to the QM subsystem along the line joining the MM  $Cl^-$  ion and the leftmost H atom in the QM subsystem.

At the first glance, the field from the full electronic density (solid red line) seems to agree rather well with the field from the multipole approximation (dashed red line), up to  $\approx 1.5$  Å, where the multipole expansion starts to diverge. However, this field is to a large degree cancelled out by the field of the QM core, making the relative error in the total (blue) much more pronounced. Additionally, the shoulder to the left of the  $Cl^-$  ion (a result of a small fraction of the electronic density being attracted there by the electrostatic potential dip from  $Cl^-$ 's induced dipole) cannot be well-represented by the multipole expansion and contributes to the charge penetration error. The issue is compounded by the fact that NGWF optimisation is driven by a gradient expression [cf. (22)] that, for consistency with a multipolar energy expression, has itself



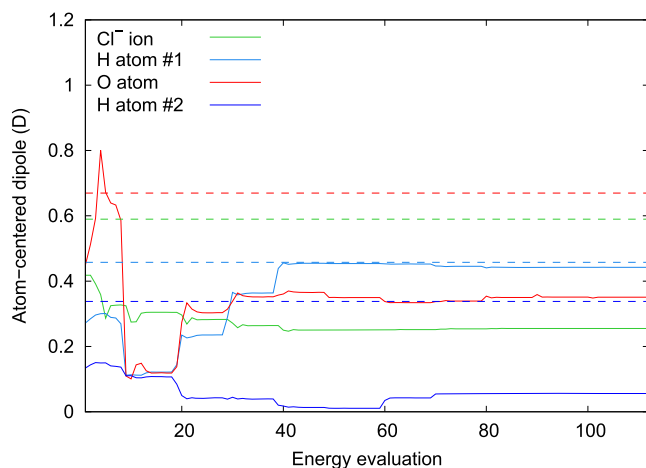
**FIG. 11.** Electric field due to the QM subsystem along the  $\text{Cl}^-$ -H line in a  $\text{H}_2\text{O}-\text{Cl}^-$  system (cf. Fig. 10). The field due to electrons alone is shown in red (solid line – full electronic density, dashed line – point-multipole approximation). The total field is shown in blue (solid line – full electronic density + core pseudopotential, dashed line – point-multipole approximation). The multipole approximation breaks down at short range (here:  $\approx 2 \text{ \AA}$ ), leading to unphysical charge transfer and lack of SCF convergence due to a polarization catastrophe.



**FIG. 13.** Electric field due to the QM subsystem along the  $\text{Cl}^-$ -H line in a  $\text{H}_2\text{O}-\text{Cl}^-$  system (cf. Fig. 10). The field due to electrons alone is shown in red (solid line – full electronic density, dashed line – point-multipole approximation). The total field is shown in blue (solid line – full electronic density + core pseudopotential, dashed line – point-multipole approximation). Once the polarization catastrophe is avoided by increased polarization damping, the multipole approximation remains accurate up to  $1 \text{ \AA}$  and SCF convergence is achieved.

been derived via an intermediate step of a multipolar expansion. In consequence, the NGWF gradient is *only* sensitive to the MM potential, field and field derivative at the QM atom centres, rather than in the entire localization sphere [cf. comment directly below (22)].

One way to avoid the QM/MM polarization catastrophe is to simply reduce the NGWF localization region (e.g., we found  $3 \text{ \AA}$  to be sufficient) – this makes the orbitals less diffuse and the point-multipole approximation more accurate.



**FIG. 12.** Atom-centered dipoles (magnitudes) on atoms in the QM subsystem (red, blue) and in the MM subsystem (green) in the course of SCF optimization for a  $\text{H}_2\text{O}-\text{Cl}^-$  system (shown in the inset of Fig. 10). With a suitable increase in polarization damping, the polarization catastrophe is avoided. Dashed lines denote reference values obtained from a fully QM calculation.

However, this would sacrifice some accuracy in the QM calculation – ONETEP calculations typically use localization radii of  $3.5 - 5 \text{ \AA}$ .

What we propose instead is to slightly attenuate QM/MM polarization interactions, leaving permanent QM/MM interactions and MM/MM polarization unchanged. We retain the Thole functional form of the damping, but we reduce Thole's  $a$  parameter by a factor of 2.45, which has the effect of attenuating the interactions at a very short range, having negligible effect elsewhere. This modification can be affected without any changes to ONETEP or TINKER simply by rescaling the apparent polarizabilities of QM atoms as seen by TINKER by a square of the above factor. The value of 2.45 has been found by numerical experiments on several small QM/MM systems, and we do not claim it to be optimal (indeed for the system studied here it leads to slight underpolarization, as seen in Fig. 12).

The practicability of the proposed solution is demonstrated in Fig. 12, which shows that all dipoles now converge to reasonable, finite values, and in Fig. 13 which shows how the total QM (electronic + core) electric field is now much better approximated by point multipoles, up to well below  $1 \text{ \AA}$ . The unexpected charge transfer from QM to the left of the  $\text{Cl}^-$  ion seen in Fig. 11 disappears, owing to the dipole on  $\text{Cl}^-$  now being well-behaved.

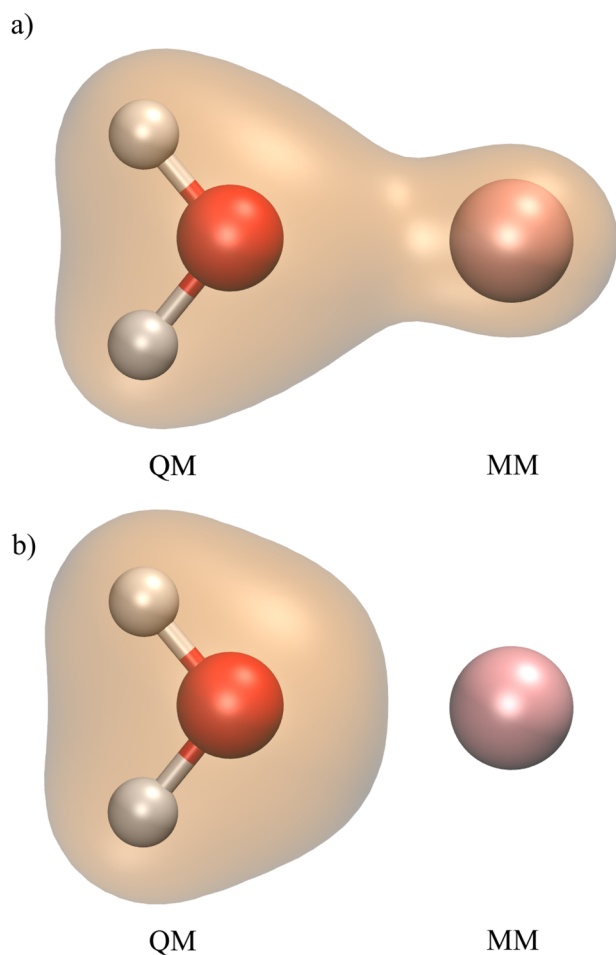
## APPENDIX B: UNPHYSICAL CHARGE TRANSFER FROM QM TO MM: DIAGNOSIS AND PREVENTION

In this section, we elaborate on the unphysical charge transfer from QM to MM that can manifest when QM/MM Pauli repulsion is not adequately taken into account. We will demonstrate that our improved model (cf. Sec. III C) addresses

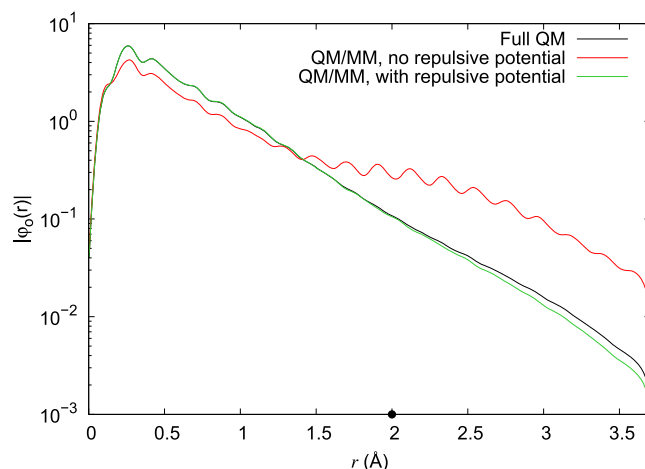
this issue satisfactorily. We use an  $\text{H}_2\text{O}:\text{K}^+$  system as an example.

Figure 14, panel (a) shows an isosurface of the electronic density at one point in the SCF optimization, where the unphysical charge transfer is apparent. The density shown is not the converged density, as the calculation fails to converge. This is because the spilled electrons accumulate near the peripheries of the orbital localization regions, and a localized-orbital formulation of DFT that assumes the orbitals to be well-decayed by the time they can be truncated cannot cope well with this situation. This is illustrated in Fig. 15, where a radial cross section of one of the NGWFs on the QM oxygen atom (red curve) is seen to differ markedly from its counterpart in a fully QM calculation (black curve).

Once a more physically sound model is used for QM/MM Pauli repulsion interactions (what we propose in Sec. III C), the charge transfer is prevented [Fig. 14, panel (b)] and the



**FIG. 14.** Electronic density isosurface ( $0.1e/\text{\AA}^3$ ) for a water- $\text{K}^+$  system. Panel (a): unphysical charge transfer from the QM subsystem to the MM subsystem. Panel (b): Repulsive potential centered on the MM atom mimics Pauli repulsion, preventing the unphysical charge transfer.



**FIG. 15.** Radial cross section of one of the  $p$  NGWFs on the QM oxygen atom for the system in Fig. 14. The NGWF localization radius is  $7 a_0 \approx 3.7 \text{\AA}$ . In the absence of a repulsive potential, the NGWF is excessively delocalized (red curve). Adding the repulsive potential restores the correct behavior (green curve), as seen by comparing against the fully QM result (black curve). The point on the  $x$  axis indicates the position of the MM  $\text{K}^+$  ion.

orbitals reacquire the correct shape (Fig. 15, green curve). The parameters used in this demonstration were  $A_{\text{K}^+} = 230 \text{ Ha/e}$  and  $\zeta_{\text{K}^+} = 1.379 a_0^{-1}$ .

## REFERENCES

- W. F. van Gunsteren and H. J. C. Berendsen, *Angew. Chem., Int. Ed.* **29**, 992 (1990).
- O. Demerdash, L.-P. Wang, and T. Head-Gordon, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **8**, e1355 (2018).
- P. S. Nerenberg and T. Head-Gordon, *Curr. Opin. Struct. Biol.* **49**, 129 (2018), Theory and simulation—Macromolecular assemblies.
- O. Demerdash, Y. Mao, T. Liu, M. Head-Gordon, and T. Head-Gordon, *J. Chem. Phys.* **147**, 161721 (2017).
- Y. Mao, O. Demerdash, M. Head-Gordon, and T. Head-Gordon, *J. Chem. Theory Comput.* **12**, 5422 (2016).
- W. L. Jorgensen, *J. Chem. Theory Comput.* **3**, 1877 (2007).
- O. Demerdash, E.-H. Yap, and T. Head-Gordon, *Annu. Rev. Phys. Chem.* **65**, 149 (2014).
- S. W. Rick and S. J. Stuart, *Rev. Comput. Chem.* **18**, 89 (2002).
- P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang, *J. Phys.: Condens. Matter* **21**, 333102 (2009).
- G. Lamoureux, A. D. MacKerell, and B. Roux, *J. Chem. Phys.* **119**, 5185 (2003).
- D. P. Geerke and W. F. van Gunsteren, *J. Phys. Chem. B* **111**, 6425 (2007).
- A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard III, *J. Phys. Chem. A* **105**, 9396 (2001).
- D.-X. Zhao, C. Liu, F.-F. Wang, C.-Y. Yu, L.-D. Gong, S.-B. Liu, and Z.-Z. Yang, *J. Chem. Theory Comput.* **6**, 795 (2010).
- G. A. Kaminski, H. A. Stern, B. J. Berne, R. A. Friesner, Y. X. Cao, R. B. Murphy, R. Zhou, and T. A. Halgren, *J. Comput. Chem.* **23**, 1515 (2002).
- P. Ren and J. W. Ponder, *J. Comput. Chem.* **23**, 1497 (2002).
- P. Ren and J. W. Ponder, *J. Phys. Chem. B* **107**, 5933 (2003).
- P. Ren, C. Wu, and J. W. Ponder, *J. Chem. Theory Comput.* **7**, 3143 (2011).
- J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, J. R. A. DiStasio, M. Head-Gordon,

- G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, *J. Phys. Chem. B* **114**, 2549 (2010).
- <sup>19</sup>P. Cieplak, J. Caldwell, and P. Kollman, *J. Comput. Chem.* **22**, 1048 (2001).
- <sup>20</sup>P. N. Day, J. H. Jensen, M. S. Gordon, S. P. Webb, W. J. Stevens, M. Krauss, D. Garmer, H. Basch, and D. Cohen, *J. Chem. Phys.* **105**, 1968 (1996).
- <sup>21</sup>N. Gresh, G. A. Cisneros, T. A. Darden, and J.-P. Piquemal, *J. Chem. Theory Comput.* **3**, 1960 (2007).
- <sup>22</sup>A. Holt, J. Boström, G. Karlström, and R. Lindh, *J. Comput. Chem.* **31**, 1583 (2010).
- <sup>23</sup>K. Burke, *J. Chem. Phys.* **136**, 150901 (2012).
- <sup>24</sup>A. Warshel and M. Levitt, *J. Mol. Biol.* **103**, 227 (1976).
- <sup>25</sup>J. Spence, Y. Huang, and O. Sankey, *Acta Metall. Mater.* **41**, 2815 (1993).
- <sup>26</sup>X. Long, J. Nicholas, M. Guest, and R. Ornstein, *J. Mol. Struct.* **412**, 121 (1997).
- <sup>27</sup>G. A. Cisneros, J.-P. Piquemal, and T. A. Darden, *J. Phys. Chem. B* **110**, 13682 (2006).
- <sup>28</sup>J. Q. Broughton, F. F. Abraham, N. Bernstein, and E. Kaxiras, *Phys. Rev. B* **60**, 2391 (1999).
- <sup>29</sup>H. Hu, Z. Lu, J. M. Parks, S. K. Burger, and W. Yang, *J. Chem. Phys.* **128**, 034105 (2008).
- <sup>30</sup>S. Ogata, E. Lidorikis, F. Shimojo, A. Nakano, P. Vashishta, and R. K. Kalia, *Comput. Phys. Commun.* **138**, 143 (2001).
- <sup>31</sup>G. Csányi, T. Albaret, M. C. Payne, and A. De Vita, *Phys. Rev. Lett.* **93**, 175503 (2004).
- <sup>32</sup>D. Fang, R. E. Duke, and G. A. Cisneros, *J. Chem. Phys.* **143**, 044103 (2015).
- <sup>33</sup>R. Khare, S. L. Mielke, J. T. Paci, S. Zhang, R. Ballarini, G. C. Schatz, and T. Belytschko, *Phys. Rev. B* **75**, 075412 (2007).
- <sup>34</sup>F. Cui and H. Li, *J. Chem. Phys.* **138**, 174114 (2013).
- <sup>35</sup>J. Dziedzic, M. Bobrowski, and J. Rybicki, *Phys. Rev. B* **83**, 224114 (2011).
- <sup>36</sup>A. J. Sodt, Y. Mei, G. König, P. Tao, R. P. Steele, B. R. Brooks, and Y. Shao, *J. Phys. Chem. A* **119**, 1511 (2015).
- <sup>37</sup>M. W. van der Kamp and A. J. Mulholland, *Biochemistry* **52**, 2708 (2013).
- <sup>38</sup>H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.* **48**, 1198 (2009).
- <sup>39</sup>N. Bernstein, J. R. Kermode, and G. Csányi, *Rep. Prog. Phys.* **72**, 026501 (2009).
- <sup>40</sup>T. Schwabe, J. M. H. Olsen, K. Sneskov, J. Kongsted, and O. Christiansen, *J. Chem. Theory Comput.* **7**, 2209 (2011).
- <sup>41</sup>M. Aida, H. Yamataka, and M. Dupuis, *Int. J. Quantum Chem.* **77**, 199 (2000).
- <sup>42</sup>J. M. Olsen, K. Aidas, and J. Kongsted, *J. Chem. Theory Comput.* **6**, 3721 (2010).
- <sup>43</sup>C. Curutchet, A. Muñoz-Losa, S. Monti, J. Kongsted, G. D. Scholes, and B. Mennucci, *J. Chem. Theory Comput.* **5**, 1838 (2009).
- <sup>44</sup>S. Caprasecca, S. Jurinovich, L. Viani, C. Curutchet, and B. Mennucci, *J. Chem. Theory Comput.* **10**, 1588 (2014).
- <sup>45</sup>C. B. Nielsen, O. Christiansen, K. V. Mikkelsen, and J. Kongsted, *J. Chem. Phys.* **126**, 154112 (2007).
- <sup>46</sup>K. Sneskov, T. Schwabe, O. Christiansen, and J. Kongsted, *Phys. Chem. Chem. Phys.* **13**, 18551 (2011).
- <sup>47</sup>S. Caprasecca, S. Jurinovich, L. Lagardère, B. Stamm, and F. Lipparini, *J. Chem. Theory Comput.* **11**, 694 (2015).
- <sup>48</sup>E. G. Kratz, A. R. Walker, L. Lagardère, F. Lipparini, J.-P. Piquemal, and G. Andrés Cisneros, *J. Comput. Chem.* **37**, 1019 (2016).
- <sup>49</sup>S. Caprasecca, C. Curutchet, and B. Mennucci, *J. Chem. Theory Comput.* **8**, 4462 (2012).
- <sup>50</sup>N. M. Thellamurege, D. Si, F. Cui, H. Zhu, R. Lai, and H. Li, *J. Comput. Chem.* **34**, 2816 (2013).
- <sup>51</sup>J. Dziedzic, Y. Mao, Y. Shao, J. Ponder, T. Head-Gordon, M. Head-Gordon, and C.-K. Skylaris, *J. Chem. Phys.* **145**, 124106 (2016).
- <sup>52</sup>Y. Mao, Y. Shao, J. Dziedzic, C.-K. Skylaris, T. Head-Gordon, and M. Head-Gordon, *J. Chem. Theory Comput.* **13**, 1963 (2017).
- <sup>53</sup>M. Schwörer, C. Wichmann, and P. Tavan, *J. Chem. Phys.* **144**, 114504 (2016).
- <sup>54</sup>D. Loco, E. Polack, S. Caprasecca, L. Lagardère, F. Lipparini, J.-P. Piquemal, and B. Mennucci, *J. Chem. Theory Comput.* **12**, 3654 (2016).
- <sup>55</sup>D. Loco, L. Lagardère, S. Caprasecca, F. Lipparini, B. Mennucci, and J.-P. Piquemal, *J. Chem. Theory Comput.* **13**, 4025 (2017).
- <sup>56</sup>R. A. Bryce, R. Buesnel, I. H. Hillier, and N. A. Burton, *Chem. Phys. Lett.* **279**, 367 (1997).
- <sup>57</sup>F. Lipparini, C. Cappelli, and V. Barone, *J. Chem. Theory Comput.* **8**, 4153 (2012).
- <sup>58</sup>F. Lipparini, C. Cappelli, and V. Barone, *J. Chem. Phys.* **138**, 234108 (2013).
- <sup>59</sup>I. Carnimeo, C. Cappelli, and V. Barone, *J. Comput. Chem.* **36**, 2271 (2015).
- <sup>60</sup>E. Boulanger and W. Thiel, *J. Chem. Theory Comput.* **10**, 1795 (2014).
- <sup>61</sup>E. Boulanger and W. Thiel, *J. Chem. Theory Comput.* **8**, 4527 (2012).
- <sup>62</sup>X. Pan, E. Rosta, and Y. Shao, *Molecules* **23**, 2500 (2018).
- <sup>63</sup>A. H. Steindal, K. Ruud, L. Frediani, K. Aidas, and J. Kongsted, *J. Phys. Chem. B* **115**, 3027 (2011).
- <sup>64</sup>See <http://www.onetep.org> for ONETEP website; accessed 08 January 2019.
- <sup>65</sup>C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, *J. Chem. Phys.* **122**, 084119 (2005).
- <sup>66</sup>C. Zhang, C. Lu, Z. Jing, C. Wu, J.-P. Piquemal, J. W. Ponder, and P. Ren, *J. Chem. Theory Comput.* **14**(4), 2084 (2018).
- <sup>67</sup>W. Kohn, *Phys. Rev. Lett.* **76**, 3168 (1996).
- <sup>68</sup>C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, *J. Phys.: Condens. Matter* **17**, 5757 (2005).
- <sup>69</sup>C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, and M. C. Payne, *Phys. Rev. B* **66**, 035119 (2002).
- <sup>70</sup>Q. Hill and C.-K. Skylaris, *Proc. R. Soc. A* **465**, 669 (2009).
- <sup>71</sup>M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, and E. Kaxiras, *J. Chem. Phys.* **114**, 5149 (2001).
- <sup>72</sup>B. Thole, *Chem. Phys.* **59**, 341 (1981).
- <sup>73</sup>P. T. van Duijnen and M. Swart, *J. Phys. Chem. A* **102**, 2399 (1998).
- <sup>74</sup>C. J. Burnham, J. Li, S. S. Xantheas, and M. Leslie, *J. Chem. Phys.* **110**, 4566 (1999).
- <sup>75</sup>T. Halgren, *J. Am. Chem. Soc.* **114**, 7827 (1992).
- <sup>76</sup>A. A. Mostofi, *On Linear-Scaling Methods for Quantum Mechanical First-Principles Calculations* (University of Cambridge, 2004).
- <sup>77</sup>Q. Hill, "Development of more accurate computational methods within linear-scaling density functional theory," Ph.D. thesis, University of Southampton, Southampton, United Kingdom, 2010.
- <sup>78</sup>J. Sala, E. Guàrdia, and M. Masia, *J. Chem. Phys.* **133**, 234101 (2010).
- <sup>79</sup>A. Stone, *Chem. Phys. Lett.* **83**, 233 (1981).
- <sup>80</sup>A. Stone and M. Alderton, *Mol. Phys.* **56**, 1047 (1985).
- <sup>81</sup>Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, *J. Chem. Theory Comput.* **9**, 4046 (2013).
- <sup>82</sup>A. J. Stone, *J. Chem. Theory Comput.* **1**, 1128 (2005).
- <sup>83</sup>A. J. Misquitta, A. J. Stone, and F. Fazeli, *J. Chem. Theory Comput.* **10**, 5405 (2014).
- <sup>84</sup>E. R. Kuechler, T. J. Giese, and D. M. York, *J. Chem. Phys.* **143**, 234111 (2015).
- <sup>85</sup>I. V. Yudanov, V. A. Nasluzov, K. M. Neyman, and N. Rösch, *Int. J. Quantum Chem.* **65**, 975 (1997).
- <sup>86</sup>J.-P. Piquemal, G. A. Cisneros, P. Reinhardt, N. Gresh, and T. A. Darden, *J. Chem. Phys.* **124**, 104101 (2006).
- <sup>87</sup>R. J. Wheatley and S. L. Price, *Mol. Phys.* **69**, 507 (1990).
- <sup>88</sup>J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- <sup>89</sup>N. Mardirossian and M. Head-Gordon, *Phys. Chem. Chem. Phys.* **16**, 9904 (2014).
- <sup>90</sup>M. J. Gillan, D. Alfé, and A. Michaelides, *J. Chem. Phys.* **144**, 130901 (2016).
- <sup>91</sup>L.-C. Lin, A. P. Seitsonen, I. Tavernelli, and U. Rothlisberger, *J. Chem. Theory Comput.* **8**, 3902 (2012).
- <sup>92</sup>R. A. DiStasio, B. Santra, Z. Li, X. Wu, and R. Car, *J. Chem. Phys.* **141**, 084502 (2014).
- <sup>93</sup>D. L. Mobley, K. L. Wymer, N. M. Lim, and J. P. Guthrie, *J. Comput.-Aided Mol. Des.* **28**, 135 (2014).

- <sup>94</sup>J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *J. Comput. Chem.* **25**, 1157 (2004).
- <sup>95</sup>W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- <sup>96</sup>T. Fox and P. A. Kollman, *J. Phys. Chem. B* **102**, 8070 (1998).
- <sup>97</sup>I. S. Joung and T. E. Cheatham, *J. Phys. Chem. B* **112**, 9020 (2008).
- <sup>98</sup>V. Vitale, J. Dziedzic, S. M.-M. Dubois, H. Fangohr, and C.-K. Skylaris, *J. Chem. Theory Comput.* **11**, 3321 (2015).
- <sup>99</sup>P. J. Dyer and P. T. Cummings, *J. Chem. Phys.* **125**, 144519 (2006).