Electrostatically Embedded Symmetry-Adapted Perturbation Theory

Caroline S. Glick,[1, 2] Asem Alenaizan,[1, 2, a)] Daniel L. Cheney,[3] Chapin E. Cavender,[4] and C. David Sherrill[1, 2, 5, b)]
[1)] *School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA*
[2)] *Center for Computational Molecular Science and Technology, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA*
[3)] *Molecular Structure and Design, Bristol-Myers Squibb Company, P.O. Box 5400, Princeton, New Jersey 08543, USA*
[4)] *Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA*
[5)] *School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0765, USA*

(Dated: 14 January 2025)

Symmetry-adapted perturbation theory (SAPT) is an *ab initio* approach that directly computes noncovalent interaction energies in terms of electrostatics, exchange repulsion, induction/polarization, and London dispersion components. Due to its high computational scaling, routine applications of even the lowest order of SAPT are typically limited to a few hundred atoms. To address this limitation, we report here the addition of electrostatic embedding to the SAPT (EE-SAPT) and ISAPT (EE-ISAPT) methods. We illustrate the embedding scheme using water trimer as a prototype example. Then, we show that EE-SAPT/EE-ISAPT can be applied for efficiently and accurately computing noncovalent interactions in large systems, including solvated dimers and protein-ligand systems. In the latter application, particular care must be taken to properly handle the quantum mechanics/molecular mechanics boundary when it cuts covalent bonds. We investigate various schemes for handling charges near this boundary, and demonstrate which are most effective in the context of charge-embedded SAPT.

I. INTRODUCTION

Quantum mechanics/molecular mechanics (QM/MM) methods are increasingly used to address the steep computational scaling of wavefunction and density functional theories.[1,2] In QM/MM methods, the system is partitioned into a QM region which is treated accurately with *ab initio* methods, and an MM region which is treated using a classical potential. The energy of the system in QM/MM methods is written as:

$$E = E_{QM} + E_{MM} + E_{QM/MM}. \tag{1}$$

The first and second terms in Eq. (1) refer to the energies of the QM and MM subsystems, respectively, and the last term refers to the interaction energy between the QM and MM regions.

The interaction potential between the QM and MM subsystems can be approximated with varying levels of accuracy, such as the electrostatic[3] and polarizable[2] embedding schemes. Electrostatic embedding accounts for the polarization of the QM region by the MM potential, while polarizable embedding further incorporates the effect of mutual polarization by the QM and MM regions.[2] Electrostatic embedding is simpler and typically utilizes a fixed set of atom-centered partial charges such as those included in classical pair-wise force fields (e.g. CHARMM[4] and AMBER[5,6]).

In this article, we extend symmetry-adapted perturbation theory (SAPT) to allow electrostatic embedding. SAPT is a correlated wavefunction approach used for computing intermolecular interaction energies in terms of physically meaningful components (electrostatics, exchange repulsion, induction, and London dispersion).[7,8] SAPT has previously been extended to intramolecular interaction energies (ISAPT),[9] and partitioning of the interaction energies into atomic (A-SAPT)[10] and functional group (F-SAPT)[11] contributions has also been developed. However, given the computational expense of the lowest order of SAPT (SAPT0), which scales as the fifth power of the number of basis functions, it is typically limited to a few hundred atoms.

Various approaches have sought to extend the applicability of SAPT to large systems, including (a) the application of efficient numerical techniques, such as density fitting[12] and local approximations,[13,14] (b) the development of alternative SAPT-based methods, such as extended SAPT (XSAPT),[15] SAPT with empirical[16–18] and semi-empirical[19,20] dispersion, and DFT-based SAPT [SAPT(DFT) or DFT-SAPT][21,22] and (c) the creation of SAPT-based force fields[23,24] and machine learning potentials.[25–27] We show here that electrostatic embedding can expand the applicability of SAPT to very large systems, even up to a complete protein, thus facilitating the computation of SAPT energies in complex molecular environments. The method remains fundamentally a SAPT approach: it computes the quantum mechanical interaction energy, and its physical components, between two sets of atoms, which we might label A and B. However, with the extensions presented here, distant or less important atoms in A and/or B can be replaced with point charge representations. We include the effect of the point charges in monomer A/B on polarizing the orbitals of the quantum mechanical atoms of A/B, and their direct contribution to the interaction energy of system A/B with system B/A [i.e., the $E_{QM/MM}$ and E_{MM} terms of Eq. (1)]. In

[a)] Present address: Chemistry Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[b)] Electronic mail: sherrill@gatech.edu

addition, leveraging our group's previously developed ISAPT framework,[9] the computations can optionally be done in the presence of a "spectator" chemical environment (labeled C), which may contain any combination of quantum mechanical atoms or atoms represented by point charges.

In the following, we illustrate the different ways that point charges can be included in an embedded SAPT calculation, then we explain the impact that the external charges have on each SAPT term. Next, we show convergence of SAPT interaction energies as the distance between the point charges and the QM region increases using water trimer. The practical utility of the method is then demonstrated through more complex chemical systems, like solvated dimers and protein-ligand complexes. In protein-ligand systems, covalent bonds that cross the QM/MM boundary must be cut, and nine different "charge schemes" that reorganize point charges at this boundary are tested. We first evaluate the charge schemes with interaction energy calculations of several model dipeptide-ligand systems as the distance between the dipeptides and ligands increases. Then, we calculate the interaction energies of a factor Xa enzyme (2CJI[28]) with two ligands that differ by a small substitution. We calculate interaction energies using each charge scheme and analyze convergence as the number of protein atoms in the QM regions increases from 47 to 446 atoms (and sometimes up to 588 atoms) and the rest of the protein is modeled by point charges. These studies help us choose the charge scheme that is most appropriate for embedded SAPT, and we observe how interaction energies change as a larger portion of the protein is modeled with QM rather than MM.

II. THEORETICAL METHODS

A. Introduction to Symmetry-Adapted Perturbation Theory

The Hamiltonian in SAPT is defined as:

$$\hat{H} = \hat{F}_A + \hat{F}_B + \lambda\hat{V} + \xi(\hat{W}_A + \hat{W}_B), \tag{2}$$

where $\hat{F}_A$ and $\hat{F}_B$ are the Fock operators for monomers A and B, respectively, $\hat{V}$ is the intermolecular interaction operator between monomer A and monomer B, and $\hat{W}_A$ and $\hat{W}_B$ are the intramolecular fluctuation potentials for monomers A and B, respectively. Wavefunction-based SAPT is a triple perturbation theory approach, in the three perturbations $\hat{V}$, $\hat{W}_A$, and $\hat{W}_B$, and λ and ξ are the perturbation strength parameters. At the SAPT0 level, which is used in this article, intramonomer electron correlation is neglected and the intermolecular potential is expanded to second order in the perturbation. This leads to the following SAPT0 interaction energy decomposition to electrostatics, exchange, induction, and dispersion components:[29]

$$\begin{aligned} E_{\text{int}}^{\text{SAPT0}} &= E_{\text{elst}}^{(100)} + E_{\text{exch}}^{(100)} \\ &\quad + (E_{\text{ind,resp}}^{(200)} + E_{\text{exch-ind,resp}}^{(200)} + \delta E_{\text{HF}}^{[2]}) \\ &\quad + (E_{\text{disp}}^{(200)} + E_{\text{exch-disp}}^{(200)}) \\ &= E_{\text{elst}} + E_{\text{exch}} + E_{\text{ind}} + E_{\text{disp}}. \end{aligned} \tag{3}$$

The superscripts indicate the perturbation order, and are specified as $E^{(nlm)}$ where n is the order in intermolecular interaction, and l and m are the orders of intramolecular flutuation of monomers A and B, respectively. The $\delta E_{HF}^{[2]}$ term, primarily accounting for higher-order induction and thus typically grouped with induction, is defined as the difference between the counterpoise-corrected[30] Hartree-Fock interaction energy and the SAPT0 electrostatics, exchange, and induction terms:

$$\delta E_{\text{HF}}^{[2]} = E_{\text{int}}^{\text{HF}} - (E_{\text{elst}}^{(100)} + E_{\text{exch}}^{(100)} + E_{\text{ind,resp}}^{(200)} + E_{\text{exch-ind,resp}}^{(200)}). \tag{4}$$

We note that this Hartree–Fock correction is in fact infinite order (because the supermolecular Hartree–Fock energy is fully self-consistent), but we prefer to label it with a [2] superscript to indicate that the correction in Eq. (4) is meant to be used with SAPT methods that include second-order induction energies. For SAPT methods including explicit third-order induction energies, the value of the Hartree–Fock correction changes, and would be labeled $\delta E_{\text{HF}}^{[3]}$.

Refs. 10 and 11 developed atomic and functional group decompositions (the A-SAPT and F-SAPT methods) of the SAPT0 interaction energy to aid in the qualitative analysis of noncovalent interactions. Ref. 9 extended the SAPT0 method to the calculation of intramolecular interaction energies (the ISAPT method). The ISAPT framework provides more generality than traditional SAPT, and so in this work we have developed an embedding approach that works for both standard SAPT and also for ISAPT; moreover, functional group partitioning, as in F-SAPT, is also enabled.

In the ISAPT approach, a chemical system is divided into three subsystems, A, B, and C, where A and B are the interacting fragments, and C represents a common molecular environment experienced by both A and B.[9] Originally, group C was used to represent the linking atoms between groups A and B. However, the approach may still be used even if C is not chemically bonded to A or B. Thus, group C might represent, for example, a set of solvent molecules, or nearby molecules in a molecular crystal. In ISAPT, a supermolecular computation is performed on the entire chemical system (A, B, and C together). Then, the occupied molecular orbitals are localized, and assigned wholly to subsystems A, B, or C. Then molecular orbitals are obtained for A/B in the absence of B/A, but in the presence of the Hartree–Fock embedding potential of C. Using these orbitals, the standard SAPT approach is then used to allow subsystems A and B to interact with each other, yielding the standard SAPT energy terms in Eq. (3). This procedure thus allows for a study of how the A–B interaction changes if A and B are both in the presence of a chemical environment C.

As one will notice from Eqs. (3) or (4), the SAPT0 interaction energy involves computation of the Hartree–Fock interaction energy. For a dimer, the interaction energy is just the difference between the dimer energy and the sum of the monomer energies (evaluated in the dimer basis to allow for counterpoise correction):[30]

$$E_{\rm int}({\rm AB}) = E_{AB} - E_A - E_B. \quad (5)$$

However, the ISAPT approach, which includes a "spectator" chemical environment C, utilizes a modified definition.[9] In ISAPT, we start with a Hartree–Fock supermolecular computation on system ABC (yielding E_{ABC}), and then we localize the occupied molecular orbitals (using the intrinsic bond orbital method of Knizia).[31] We identify the local occupied orbitals associated with subsystem C, and compute the Hartree–Fock energy associated with those local orbitals and the nuclei associated with subsystem C, $E_C^{\rm LO}$, where the superscript LO denotes local orbitals. Those subsystem C local orbitals and nuclei also define the Hartree–Fock embedding potential of subsystem C, and we use this fixed embedding potential to compute the Hartree–Fock energies of subsystems A and B in this potential, which we might denote $E_{A[C]}$ and $E_{B[C]}$, respectively (where the brackets indicate the embedding potential).

With these definitions for the energies, ISAPT defines a Hartree–Fock interaction energy between monomers A and B in the embedding environment of C as:[9]

$$E_{\rm int}^{\rm HF}({\rm AB,ISAPT}) = E_{ABC} - E_{A[C]} - E_{B[C]} + E_C^{\rm LO}, \quad (6)$$

where each term represents a Hartree–Fock computation on the designated system or subsystem, as defined above. For counterpoise correction, all computations are performed using the union of basis functions on subsystems A, B, and C. Eq. (6) is just a generalization of Eq. (5) in which subsystem C is always present (first three terms), and then the internal energy of C, $E_C^{\rm LO}$, has to be added to the expression so that it makes no net contribution to the interaction of A and B (it contributes positively in E_{ABC} and negatively in both $E_{A[C]}$ and $E_{B[C]}$).

While this approach seemed suitable to us in the original context of computing the interaction energy between two groups A and B within a single molecule, connected by a linker C, we do note that it has a drawback if we then apply the ISAPT methodology to systems in which C is not covalently connected to A and B. With these definitions, for a non-covalent trimer, if we compute the ISAPT Hartree–Fock interaction energies for each dimer, while treating the remaining monomer as a spectator group represented by a Hartree–Fock embedding potential, the sum $E_{\rm int}^{\rm HF}({\rm AB,ISAPT}) + E_{\rm int}^{\rm HF}({\rm AC,ISAPT}) + E_{\rm int}^{\rm HF}({\rm BC,ISAPT})$ does not add up to the overall Hartree–Fock interaction energy of the trimer. For the water trimer example discussed below, the sum of the ISAPT HF interaction energies, including $\delta E_{\rm HF}^{[2]}$ corrections computed via Eq. (6), is -13.50 kcal mol^{-1}, vs. a HF interaction energy of -11.26 kcal mol^{-1}. (Interestingly, the sum of the ISAPT interaction energies without dispersion or $\delta E_{\rm HF}^{[2]}$ corrections is closer to the HF interaction energy in this case, -10.66 kcal mol^{-1}.) Our group is considering alternative definitions for $E_{\rm int}^{\rm HF}({\rm AB,ISAPT})$ that might be more suitable for $\delta E_{\rm HF}^{[2]}$-corrected ISAPT applied to molecular clusters instead of single molecules.

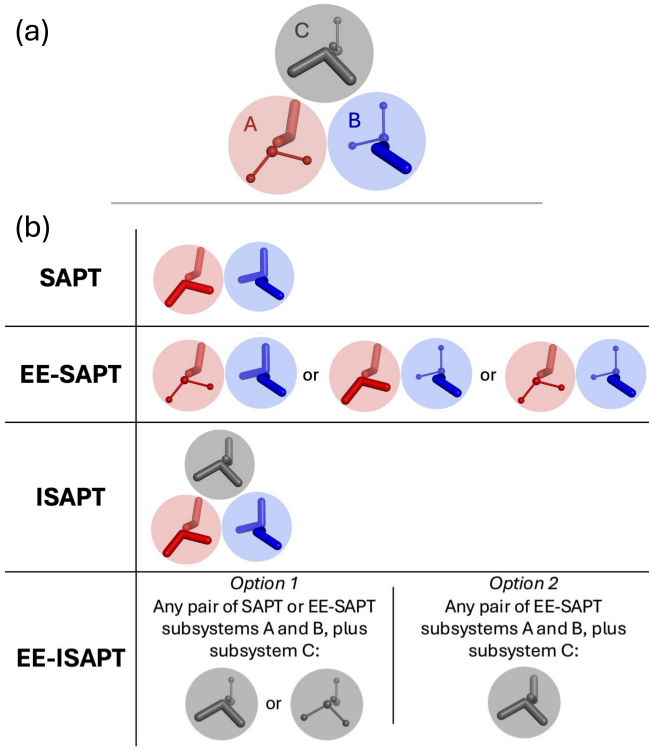

FIG. 1. A water hexamer illustrating possible scenarios for including point charges in the SAPT and ISAPT procedures. (a) Subsystems A and B are the interacting fragments while optional subsystem C is the environment. Water molecules drawn in ball-and-stick representation are selected for possible substitution with point charges. (b) Table of the possible ways to include point charges (or not) for the water hexamer in (a). Rows SAPT and ISAPT show the conventional system set-up for these methods, in which every atom in represented quantum mechanically. The other rows show different possibilities for including point charges in SAPT and ISAPT, named EE-SAPT and EE-ISAPT respectively.

Below we discuss our approach for adding the effect of point charges to fragments A, and/or B, and/or C (if present), to provide a general and flexible approach to electrostatic embedding in SAPT0.

B. Charge embedding schemes available in EE-(I)SAPT

Because SAPT0 computes the interaction energy between subsystems A and B, possibly in the presence of an optional spectator environment C (if the ISAPT formalism is used), there are several possible ways for including point charges in a given SAPT0 calculation using this three-subsystem (A, B, C) framework. We illustrate these possible schemes in Figure 1 using the geometry of water hexamer in the prism configuration.[32]

As shown in Figure 1 (a), the water hexamer is divided into three subsystems (A, B, and C), each consisting of two water molecules. In general, subsystem C (if it is present) may or may not contain explicit QM atoms, while subsystems A and B must necessarily contain QM atoms in the SAPT procedure. Figure 1 (b) illustrates the possible scenarios for the inclusion of point charges in A and/or B without a subsystem C (EE-SAPT), as well as possibilities for the inclusion of a monomer C (EE-ISAPT). Point charges can be included in any combination of A, B, C (or none). The color scheme in Figure 1 (b) indicates whether a given subsystem explicitly contributes to the interaction energy. Since subsystem C only indirectly modulates the interaction energy between A and B, the C atoms (including the point charges) do not explicitly contribute to the A-B interaction energy and hence are drawn in gray. By contrast, the A and B atoms explicitly participate in the interaction and therefore are colored.

Choosing a specific partition scheme depends on the target problem. For example, point charges in C can approximately represent solvent molecules that modulate the SAPT0 interaction energy, while point charges in A and B can replace far portions of fragments A or B (e.g. distant protein atoms in computations of protein-ligand interactions). We illustrate these possibilities below in the Results and Discussion, focusing on solvated dimers and protein-ligand systems.

C. Physical effects of embedding charges

Here we discuss how an embedding point charge will affect the chemical system and the interaction between fragments A and B. We focus initially on the effect of embedding charges in subsystems A and B, and then later consider the effect of embedding charges in the environmental group C.

1. Embedding charge-point charge interactions

First, an embedding point charge will experience a Coulomb's law attraction or repulsion with each of the nuclei in the system, and with each of the other embedding charges. Because we are ultimately interested in computing an interaction energy between fragments A and B, we have some flexibility in how we account for some of these interactions. For example, if some of the atoms in fragments A and B are represented by embedding charges, then it is mandatory to include the charge-charge interactions between each embedding charge in A and each embedding charge in B:

$$E_{\text{elst}}^{\text{A-extern,B-extern}} = \sum_{a\in A}\sum_{b\in B}\frac{z_a z_b}{R_{ab}}, \tag{7}$$

where we will use lowercase z to denote embedding charges (as opposed to uppercase Z to denote nuclear charges). In our bookkeeping scheme, we will classify this term as a separate electrostatic term not included in $E_{\text{elst}}^{(100)}$. However, with respect to the interaction energy, it is not necessary to compute Coulomb interactions between all pairs of embedding charges *within* fragment A (or B), because these would simply cancel out when computing the electrostatic interaction between fragments A and B. For this reason, we only compute the relevant Coulomb interactions between embedding charges, i.e., those in Eq. (7).

Interactions between external charges in A and nuclei in B, and between external charges in B and nuclei of A, also contribute to the electrostatic part of the interaction energy:

$$E_{\text{elst}}^{\text{extern,nucl}} = \sum_{a\in A}\sum_{b\in B}\left(\frac{z_a Z_b}{R_{ab}} + \frac{Z_a z_b}{R_{ab}}\right). \tag{8}$$

This contribution is added to $E_{\text{elst}}^{(100)}$.

2. Embedding charge-electron interactions

Next, we consider how embedding charges will interact with the system's electrons. There are both direct and indirect effects. The direct effect is the interaction of the electrons of A or B with the embedding charges of B or A. The potential generated by the embedding charges in subsystem X (where X is A, B, or C) may be written as

$$V^{\text{X-extern}}(\vec{r}_1) = \sum_{x\in X}\frac{z_x}{r_{1x}}, \tag{9}$$

where $\{z_x\}$ are the embedding charges assigned to fragment X. In the given atomic orbital basis set, the matrix representation of this potential is just

$$V_{\mu\nu}^{\text{X-extern}} = \int_{\mathbb{R}^3} d^3 r_1 \phi_\mu(\vec{r}_1)\left(\sum_{x\in X}\frac{z_x}{r_{1x}}\right)\phi_\nu(\vec{r}_1), \tag{10}$$

where ϕ_μ and ϕ_ν are two basis functions.

The electrostatic interaction between the electrons of A/B and the embedding charges of B/A is then just

$$E_{\text{elst}}^{\text{extern,elec}} = 2\sum_{\mu\nu} D_{\mu\nu}^A V_{\mu\nu}^{\text{B-extern}} + 2\sum_{\mu\nu} D_{\mu\nu}^B V_{\mu\nu}^{\text{A-extern}}, \tag{11}$$

where the density matrix for subsystem X is defined by

$$D_{\mu\nu}^X = \sum_i^X C_{\mu i}^X C_{\nu i}^X, \tag{12}$$

with the summation running over occupied molecular orbitals of fragment X. $C_{\mu i}^X$ are the molecular orbital coefficients of fragment X, and the factors of 2 in Eq. (11) account for doubly-occupied molecular orbitals in the Restricted Hartree–Fock formalism. The term $E_{\text{elst}}^{\text{extern,elec}}$ is added to $E_{\text{elst}}^{(100)}$.

The embedding charges will also contribute to the induction terms. Induction is computed by coupled-perturbed Hartree–Fock (CPHF), and quantifies the stabilization due to the polarization of orbitals in A/B in response to the potential field of the embedding charges in B/A. This is accounted for by simply adding $V^{\text{A-extern}}$ to the electrostatic potential of A when performing the CPHF computation on monomer B, and vice versa. These effects are included in the $E_{\text{ind,resp}}^{(200)}$ term.

Embedding charges also have indirect effects on the interaction energy. Embedding charges in A will affect the molecular orbitals of A, and likewise for embedding charges in B. (Embedding charges or QM atoms in the environmental group C will likewise affect orbitals in A and B). The modified density in a monomer will cause changes to all of its interaction energy terms.

3. Hartree–Fock correction term

In low-order SAPT, it is common to account for higher-order induction by applying the Hartree–Fock correction of Eq. (4), which involves the Hartree–Fock interaction energy $E_{\text{int}}^{\text{HF}}$.

For counterpoise correction, we compute all terms in the dimer basis. When computing these Hartree–Fock energies, for simplicity we include *all* Coulomb interactions between embedding charges and nuclei into the nuclear repulsion energies, even though many of these contributions cancel out in determining $E_{\text{int}}^{\text{HF}}$ (e.g., interactions between nuclei of A and embedding charges of A). Other contributions (e.g., interactions between nuclei of A and embedding charges of B) do not cancel in computing $E_{\text{int}}^{\text{HF}}$, but they do cancel when computing $\delta E_{\text{HF}}^{[2]}$, because these same contributions were also included in $E_{\text{elst}}^{(100)}$. No charge-charge interactions involving embedding charges contribute to $\delta E_{\text{HF}}^{[2]}$.

We also include $V^{\text{X-extern}}$ in Hartree–Fock computations of monomer X, as well as Hartree–Fock computations of any dimer or trimer involving subsystem X (e.g., as might be needed to compute the Hartree–Fock interaction energy in Eqs. (5) or (6)). This is necessary to allow the orbitals of any monomer, dimer, or trimer including subsystem X to respond to the embedding charges. Of course this then affects both $E_{\text{int}}^{\text{HF}}$ and $\delta E_{\text{HF}}^{[2]}$ (which is computed from $E_{\text{int}}^{\text{HF}}$).

4. Embedding charges in environment C

Embedding point charges (and any embedding Hartree–Fock potential from QM atoms) in the common environment C are largely "spectators" with respect to the interaction between A and B, and do not contribute explicitly to any of the A–B interaction energy SAPT terms. Of course, embedding charges or QM atoms in C do affect the orbitals of A and B, because in ISAPT the orbitals of A and B are determined in the presence of the fixed potential of environment C,[9] and the modification of the orbitals thus indirectly affects all SAPT terms for the A–B interaction. The addition of embedding charges in C presents no extra difficulty for the ISAPT procedure. Originally, the orbitals of A and B would be determined in the presence of the Hartree-Fock potential V^C for any QM atoms in C (this term is added to the Fock operator). Now, we also add $V^{\text{C-extern}}$ for any embedding charges in C.

In the presence of a group C, the Hartree–Fock interaction energy $E_{\text{int}}^{\text{HF}}$ is defined as in Eq. (6). We include $V^C + V^{\text{C-extern}}$ for all Hartree–Fock computations involving group C. In our implementation, for simplicity we include nuclear / embedding charge interactions when computing nuclear repulsion energies in Hartree–Fock computations required for $\delta E_{\text{HF}}^{[2]}$. However, these contributions all cancel when computing $E_{\text{int}}^{\text{HF}}$. Interactions between embedding charges in C and embedding charges in A or B would cancel out in computing $E_{\text{int}}^{\text{HF}}$ and are not included in the nuclear repulsion energies.

Potentials generated by environment C affect $\delta E_{\text{HF}}^{[2]}$ indirectly, just like the SAPT terms, by the changing of the monomer A and monomer B orbitals due to the presence of the potential in C. There is also a somewhat subtle, more direct effect on $\delta E_{\text{HF}}^{[2]}$. In the case of no environment C, the density ascribed to monomer A in the dimer AB is not identical to the density of isolated monomer A (this is the polarization effect whose energetic contribution is captured in $E_{\text{ind,resp}}^{(200)}[A \leftarrow B]$). In the case of an environment C, again the density associated with monomer A in the Hartree–Fock computation for ABC is different than that for AC. However, $\delta E_{\text{HF}}^{[2]}$ captures not only how this density change interacts with monomer B, but also how it interacts with environment C (and, likewise, how the change in the density of monomer B interacts with environment C). This might be considered a contribution to the "effective" A–B interaction in the environment C. This effect was captured in the original ISAPT for quantum mechanical environments C, and here it is also captured for point charges in C.

5. F-SAPT analysis for the embedding charges

The addition of embedding charges does not impede partitioning the SAPT components into contributions between pairs of functional groups via the F-SAPT approach.[11] For example, the electrostatic interaction of the embedding charges of fragment A with fragment B may be written:

$$E_{\text{elst}}^{\text{A-extern, B}} = \sum_{\bar{b}\in B} \int_{\mathbb{R}^3} d^3 r_1 \phi_{\bar{b}}(\vec{r}_1)\phi_{\bar{b}}(\vec{r}_1) \times \left(\sum_a \frac{z_a}{r_{1a}} + \sum_{aB} \frac{z_a Z_B}{R_{aB}} + \sum_{ab} \frac{z_a z_b}{R_{ab}} \right), \quad (13)$$

where $\phi_{\bar{b}}$ are the occupied orbitals corresponding to fragment B, z_a are the embedding charges associated with fragment A, Z_B are the nuclear charges in fragment B, and z_b are the embedding charges in fragment B. If we use local orbitals $\phi_{\bar{b}}$ as in the F-SAPT and ISAPT procedures, then these interactions can be trivially assigned to particular functional groups as usual in F-SAPT.[11]

The induction contribution arising from the point charges can also be separated. As detailed previously,[10] the induction energy in SAPT theory may be partitioned into contributions from pairs of functional groups. The induction energy from the polarization of monomer B due to the electric field of monomer A is broken down into pair contributions, where each pair involves one local source of the electric field of A (a nucleus or an electron in a local orbital), and one local orbital

in B whose electrons will be excited in response to this field contribution. This same approach works in the present case, because the embedding charges in A make their own contributions to the overall potential due to monomer A, and can be handled analogously to the nuclei of monomer A (and vice versa for monomer B).

This separability enables the use of the standard F-SAPT analysis and visualization tools to be extended to the analysis of the embedding charges. For example, Figure 1 shows the application of the F-SAPT order-1 analysis for the visualization of the interaction between water molecules. The energetic contributions of the embedding charges to the SAPT0 interaction energy and the electrostatics and induction terms are also available. The embedding extensions of F-SAPT and ISAPT described here have been added to the open-source quantum chemistry program Psi4 and are avialable in the current release.[33]

III. RESULTS AND DISCUSSION

First, we use water trimer to demonstrate the embedding method and its convergence at long range. Then, we illustrate more practical applications of electrostatically embedded SAPT: calculating interaction energies of solvated dimers and protein-ligand systems.

A. The water trimer: A prototype example

We first illustrate the embedding procedure and various partitioning schemes using water trimer as a prototype example. Figure 2 (a) shows the water trimer geometry where one molecule is assigned to fragment A and one is assigned to fragment B. The third, shaded molecule is either grouped with A or B (SAPT) or assigned to the external environment C (ISAPT). Additionally, the shaded water molecule is either treated quantum mechanically or substituted by TIP3P[34] point charges. TIP3P was chosen for simplicity, but we have repeated the experiment using Restrained Electrostatic Potential[35,36] (RESP) and Minimal Basis Iterative Stockholder[37] (MBIS) charges instead of TIP3P. The results are presented in the Supplementary Material; using RESP and MBIS rather than TIP3P did not change the overall conclusions. The shaded water molecule is displaced from equilibrium along the positive y-axis and then the SAPT0/jun-cc-pVDZ interaction energy is computed at the various configurations.

Figure 2 (b) shows the SAPT0 interaction energy and compares the electrostatic embedding approximation with the full quantum mechanical treatment. As expected, the effect of the shaded molecule on the interaction energy gradually decreases as the molecule is moved away from the the rest of the trimer. All the interaction energies approach the dimer interaction energy (-4.87 kcal mol^{-1}) as the displacement increases to 10 Å. At short range, the various partitioning schemes leads to varying interaction energies, and the effect of electrostatic embedding becomes more apparent. Tables S-1–S-6 in the Supplementary Material show the detailed SAPT0 interaction energy and its decomposition.

When the shaded molecule is assigned to the environment C and treated quantum mechanically, it polarizes the two other molecules but does not participate directly in the interaction. The dimer interaction becomes more attractive (the interaction energy becomes more negative) by 1.34 kcal mol^{-1} at the equilibrium distance compared to the isolated dimer, shown by the brown solid line in Fig. 2 (b). Electrostatic embedding underestimates the polarization effect by 0.39 kcal mol^{-1} at equilibrium but quickly agrees with the full QM calculation as the shaded molecule is displaced away. [See the brown dashed line in Fig. 2 (b).] The error is largely in the electrostatic component of the SAPT0 energy as shown in Table S-1 and S-2.

When the shaded molecule is grouped with either A or B [turquoise and purple in Fig. 2 (b), respectively], it directly participates in the interaction. Therefore, the SAPT0 interaction energy increases by over a factor of 2 compared to the isolated dimer energy. At the equilibrium distance, electrostatic embedding overestimates the interaction energy by $1.0-1.2$ kcal mol^{-1} compared to the full quantum mechanical treatment. However, both the absolute and relative errors decrease as the displacement increases.

The inset in Figure 2 (b) shows the effect of electrostatic embedding on the SAPT0 interaction energy components when the shaded molecule is grouped with A or B. At the equilibrium distance, all SAPT0 components display substantial errors, especially the exchange repulsion. Large errors should be expected at small intermolecular distances since electrostatic embedding does not account for certain interactions, namely exchange, dispersion, parts of the induction, and charge-penetration electrostatics.[39] Fortunately, however, the positive errors in the electrostatics, induction, and dispersion terms are mostly cancelled by the negative error in the exchange term, so that the error in the total interaction energy is smaller in magnitude than the errors in the individual components. The errors in the energy components rapidly disappear as the displacement increases, in part due to the lower magnitude of the interaction, but primarily because charge embedding is more appropriate at longer displacements. In addition, the favorable error cancellation between energy components persists at larger distances. Tables S-3–S-6 show the detailed SAPT0 component data.

B. Solvated systems with EE-ISAPT

The water trimer example illustrates the SAPT0 electrostatic embedding procedure, but it is not a realistic example of the kind of problems to which the procedure would actually be applied. Here, we show the utility of the method for calculating intermolecular interactions of solvated systems.

The two solvated systems studied are the pyridinium:benzene dimer (Figure 3) and the benzene dimer, both in water. The geometries are those used in our group's previous study of water's polarization tuning of π-π interactions.[40] In summary, the tilted T-shaped pyridine:benzene of the

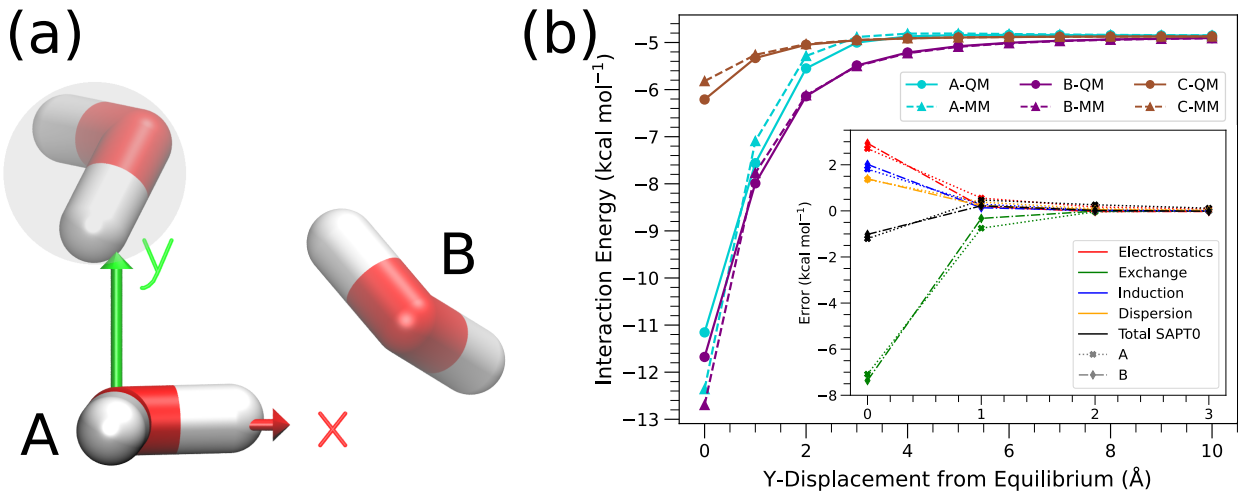

FIG. 2. The water trimer geometry and interaction energy. (a) The water trimer at equilibrium. The shaded water molecule is treated as either a QM or a TIP3P (MM point charge) molecule. It can be grouped with A or B or it can define a spectator chemical environment, subsystem C in the ISAPT framework. The energy is computed at various configurations where the shaded molecule is displaced from the equilibrium geometry along the y-axis. The equilibrium geometry is taken from Ref. 38. (b) The SAPT0/jun-cc-pVDZ interaction energy for the water trimer at various displacements and partitions of the shaded water molecule. The legend X-QM/MM means that the shaded water is considered a part of subsystem X, and it is treated via QM or MM. For example, "A-QM" adds the shaded water together with water A and treats this two-water subsystem-A fully self-consistently with QM; its interaction energy with QM water B is computed using normal SAPT. "C-QM" represents an ISAPT computation in which the shaded water is treated as a HF embedding potential, and "C-MM" represents an EE-ISAPT computation in which the shaded water is treated as a point-charge embedding potential. The inset displays the error in the SAPT0 components when the shaded molecule, grouped with A (x) or with B (diamond), is substituted by TIP3P point charges. The x-axis remains y-displacement from equilibrium.

S66 dataset[41,42] was functionalized then optimized with B3LYP-D3M(BJ)/aug-cc-pVDZ. The dimer geometries were then fixed, and molecular dynamics simulations were used to determine locations of the solvating water molecules. Specifically, the snapshots used in this study are named 'HYD8-7m2' and 'HYD8-3m1' in Ref. 40. To generate reference interaction energies, systems were chosen such that a full QM calculation could be run in a reasonable amount of time (about one day). The maximum system size of the solvated pyridinium:benzene studied has 136 water molecules that contribute to a full system size of 432 atoms (148 heavy). For the benzene dimer, the maximum system size has 118 waters. Each water included has a closest contact that is less than 7 Å from the dimer. These waters are considered monomer C in the ISAPT calculations, thus they are indirectly affecting the interaction energy between the two solute monomers.

For the pyridinium:benzene system, SAPT0 with a jun-cc-pVDZ basis set calculates a gas phase interaction energy of -8.97 kcal mol^{-1} in 19 seconds on six cores of an Intel i9-9820X processor. When the full environment of waters is added and represented with quantum mechanics, the interaction energy decreases in magnitude to -4.79 kcal mol^{-1}, but the wall time is 33 hours and 40 minutes with the same computational resources. To reduce this computational cost, we replace all waters with TIP3P charges: 0.417 a.u. for H and -0.834 a.u. for O. This incurs an error of –0.35 kcal mol^{-1} relative to our reference. An interaction energy of -4.99 kcal mol^{-1}, within 1 kJ mol^{-1} of the reference, can be calculated in just 4.5 minutes after replacing the outermost 118 waters with point charges. For this claculation, the only waters included quantum mechanically were those with closest contacts less than 3 Å, totaling 18 water molecules. As more waters are included (with closest contacts less than 4 Å, 5 Å, and 6 Å), errors relative to the reference value continue to decrease, but the computational cost increases, shown in Figure 3.

The interaction energy components, also shown in Fig. 3, reveal that exchange and induction energies are generally insensitive to the water molecules being modeled with MM rather than QM. For this test case, dispersion is not sufficiently attractive relative to the reference value (because it is not captured by the electrostatic embedding), and the error is above 1 kJ mol^{-1} when all waters are modeled with MM. This error is quickly reduced below 1 kJ mol^{-1} when 18 waters are modeled with QM rather than TIP3P charges. Electrostatics takes the longest to converge, and its error is larger than 1 kJ mol^{-1} with 18 QM waters. Due to error cancellation, the interaction energy converges more quickly than the electrostatics term.

The benzene dimer is bound primarily by dispersion forces, and our group's previous study concludes that the solute-solute interactions of a neutral dimer are affected very little by solvent molecules.[40] The gas phase and the solvated reference (118 QM waters) interaction energies of this tilted T-shaped dimer differ only by 0.10 kcal mol^{-1}. Shown in Figure 4, modeling all 118 waters with TIP3P charges returns an error in interaction energy, relative to the reference value, of 0.27

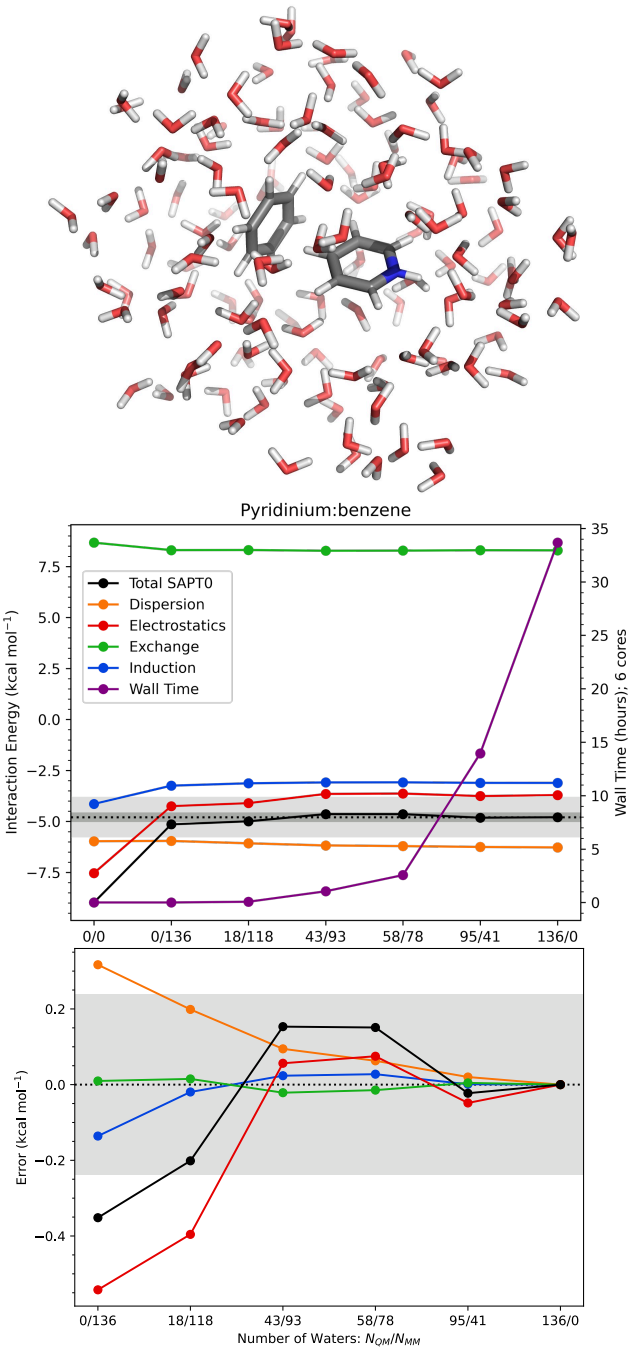

FIG. 3. The pyridinium:benzene dimer in 136 waters (top), and the SAPT0/jun-cc-pVDZ interaction energy with its energy components (middle) for the pyridinium:benzene dimer as an increasing number of the surrounding 136 waters are modeled by QM rather than TIP3P charges. The light gray box bounds the final interaction energy (with 136 QM waters) by $\pm$ 1 kcal mol^{-1}, and the dark gray bounds the final value by $\pm$ 1 kJ mol^{-1}. Six cores of an Intel i9-9820X processor were used for the timings. The bottom graph plots errors relative to the energy of the fully QM calculation, and the gray box indicates $\pm$ 1 kJ mol^{-1}.

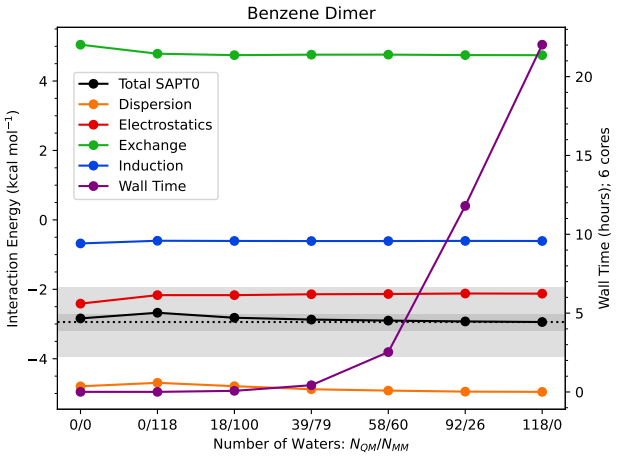

FIG. 4. The SAPT0/jun-cc-pVDZ interaction energy and components of the benzene dimer as an increasing number of the surrounding 118 waters are modeled by QM rather than TIP3P charges. The light gray box bounds the final interaction energy (with 118 QM waters) by $\pm$ 1 kcal mol^{-1}, and the dark gray bounds the final value by $\pm$ 1 kJ mol^{-1}. Six cores of an Intel i9-9820X processor were used for the timings.

kcal mol^{-1}, which is just above 1 kJ mol^{-1}. When the closest 18 waters are modeled with QM instead, the error reduces to 0.12 kcal mol^{-1}. The wall time of this calculation is 4 minutes and 14 seconds, a drastic reduction relative to the wall time of 22 hours when all 118 waters are modeled with QM.

For both dimers, we conclude that SAPT with electrostatic embedding correctly accounts for the influence of water solvent on the interaction energy. When modeling a small number of waters with QM, this method returns accurate interaction energies (within 1 kJ mol^{-1}) in just a few minutes, rather than a day.

C. Protein-ligand interactions with EE-SAPT: A detailed analysis

We now show the utility of SAPT with electrostatic embedding for systems where monomer A and/or B are too large to include fully quantum mechanically. For example, protein-ligand systems often have thousands of atoms, and so quantum mechanical studies must consider only a small portion of the protein due to the high computational expense.[43,44] Here, we present SAPT0 interaction energies of protein-ligand complexes, where part of the proteins are modeled with point charges, but still considered part of monomer B. This is inherently a more complicated task than the previous two examples for two reasons. First, including the point charges in monomer B will directly affect the interaction energy, and higher errors are expected due to embedding relative to a system where the point charges are placed in environment C (see Section III A). Second, separating a covalently-bound molecule into QM and MM regions, while minimizing additional error, is a non-trivial task.

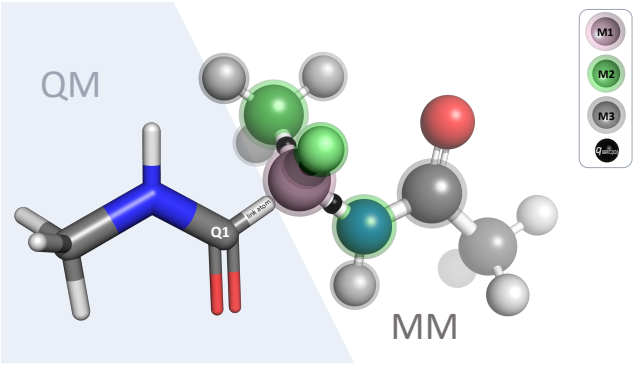

FIG. 5. A simple protein system is divided into QM (blue background) and MM (white background) regions by cutting a carbon-carbon bond. The QM region is capped by a hydrogen link (HL) atom. The carbon in the MM region that is involved in the cut bond is M1 and highlighted by the pink sphere. M2 atoms (green) are directly connected to M1, and M3 atoms (gray) are directly connected to M2 atoms. For redistributed charge schemes BRC and BRCD, charges are redistributed to the midpoints of the M1–M2 bonds, represented by the black bands.

When forming QM and MM regions within a molecule, the frontier covalent bonds connecting the QM and MM regions must be "cut." We refer to the QM and MM atoms directly connected by this frontier bond as Q1 and M1, respectively, and illustrate these in Figure 5. This cut creates an open valency on the Q1 atom, which is commonly satisfied by a hydrogen "link" (HL) atom. This hydrogen is placed along the Q1–M1 frontier bond, at a bond distance determined by a function of the force field bond stretch parameters, as proposed by Truhlar and coworkers.[45] The parameters chosen are from the same force field used to generate atomic partial charges for the MM region. Because we are modeling a protein, we use ff19SB,[6] which results in a Q1–HL bond length around 1.09–1.11 Å.

The close proximity of the MM region to the HL atom can cause overpolarization. To reduce this effect, several methods of altering charges at the boundary have been proposed in the literature. There is limited consensus on which is the best approach,[45–48] and benchmarking these different charge schemes typically considers proton affinities rather than interaction energies.[47,49,50] We have implemented nine different schemes, including elimination (Z1, Z2, Z3) and balanced (DZ1, DZ2, DZ3, BRC, BRCD, BRC2) schemes. To explain the specific schemes, labels M2 and M3 are used in addition to M1, Q1, and HL. Shown in Figure 5, M2 atoms are the atoms in the MM region directly connected to M1, and M3 atoms are atoms in the MM region directly connected to M2 atoms. Charges for each atom (i) are represented as q_i.

The charge elimination schemes, Z1,[51] Z2,[52] and Z3,[52] set charges of different MX atoms to 0, where X = 1, 2, and/or 3. In Z1, $q_{M1} = 0$; in Z2, $q_{M2} = q_{M1} = 0$; and in Z3, $q_{M3} = q_{M2} = q_{M1} = 0$. While simple, these elimination schemes change the overall charge of the MM region. We have also implemented charge balancing schemes that distribute the charge necessary to return the frontier MM residue to its original integer charge, thereby conserving the charge of the MM region. Our selected Amber force field, ff19SB,[6] assigns point charges such that each residue and endcap sums to the appropriate integer charge.

In DZ1, $q_{M1} = 0$, and the charge necessary to return the residue in the MM region to its integer charge is distributed evenly among the MM atoms in that residue. The same is true of DZ2[47] and DZ3, except $q_{M2} = q_{M1} = 0$ and $q_{M3} = q_{M2} = q_{M1} = 0$, respectively. In the balanced redistributed charge scheme (BRC),[48] $q_{M1} = 0$, and the charge needed to return the residue to an integer charge is distributed evenly to the midpoint of the M1–M2 bonds. BRC2[48] is similar, except that it redistributes the charge by adding it evenly to the charges on the M2 atoms. The balanced redistributed charge and dipole scheme (BRCD)[48] extends the BRC scheme by doubling the redistributed charge at the M1–M2 bond midpoints, then subtracting the redistributed charge from q_{M2}. This preserves the M1–M2 bond dipoles.

We first test these embedding schemes on various ligand:dipeptide complexes. Beginning with these simple systems allows us to benchmark the charge schemes against a fully QM interaction energy. The ligand is taken from the PDB entry 2CJI[28], which includes the factor Xa (fXa) enzyme, key in blood coagulation, and a ligand that targets fXa's S1-pocket. Previously, our group used F-SAPT to reveal why chloro-substituted ligands of various fXa systems bind more strongly to the negatively-charged S1-pocket than methyl-substituted variants.[43] That study truncated the ligand to contain 26 atoms for the chlorinated (Cl) variant, and 29 for the methylated (Me) variant, to include the atoms involved in the S1 binding pocket and maintain a reasonable system size. Here for simplicity we use the same truncated ligands. After some simple tests of these ligands interacting with dipeptides, later in Section III D we consider these ligands interacting with the full fXa protein.

Each dipeptide in our tests contains two of the following amino acids: glutamic acid (GLU), histidine (neutral, HIE, and protonated, HIP) and glutamine (GLN). This group of residues is chosen due to their diversity in overall charge. HIE and GLN are both neutral, GLU has a charge of –1, and HIP has a charge of +1. For each dipeptide created, the N-terminus is capped with an acetyl group (ACE), and the C-terminus is capped with N-methylamide (NME). Both of these groups are neutral.

Initially, we calculate the fully QM interaction energy of the ligand and a dipeptide containing GLN (neutral) and GLU (–1 charge), referred to as 'aceGLN(GLUnme)' and shown in Figure 6. Then, we model GLU (and the capping NME) with point charges from ff19SB, cutting between C_α of GLN (Q1) and the carbonyl carbon of GLN (M1). The point charges are altered according to each of the nine schemes. The top panel of Figure 6 reports the interaction energy errors for each charge scheme, relative to a fully QM calculation, for this system. It also shows how these errors change as the intermolecular distance increases between the ligand's chlorine and the closest point charge. In this case, BRC2 returns an especially accurate interaction energy relative to QM. At 6.7 Å, the QM interaction energy is –9.45 kJ mol^{-1} and this energy is –9.62

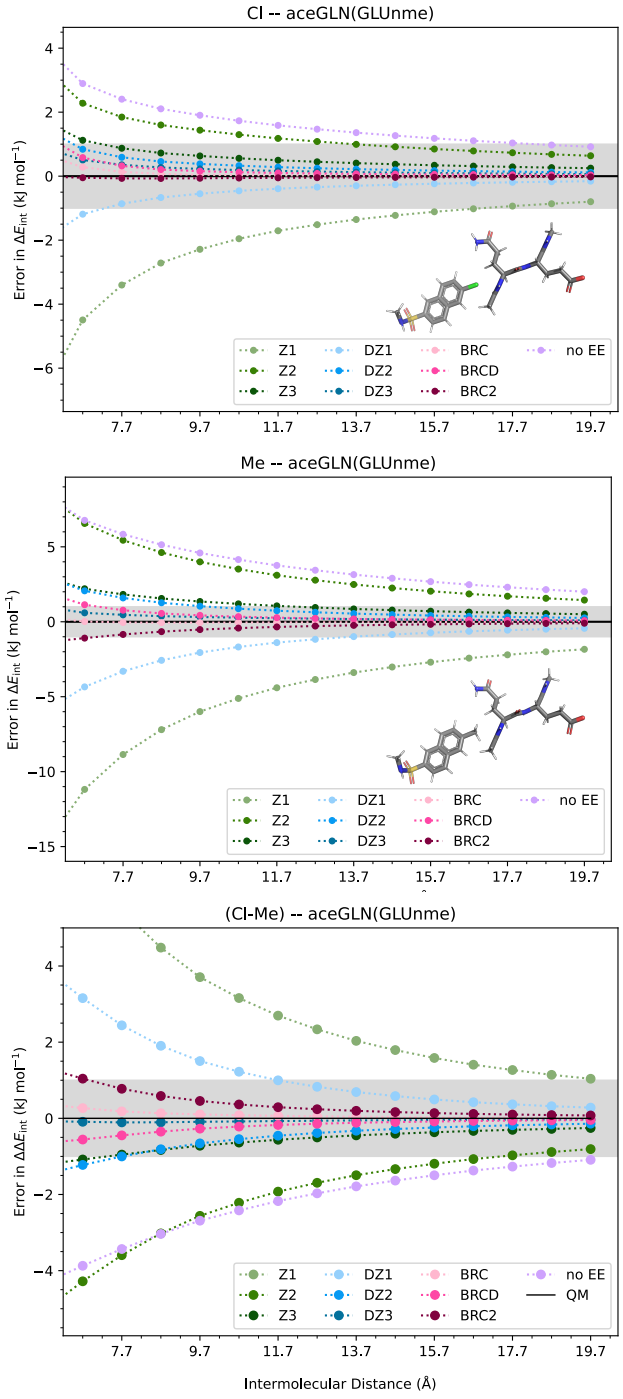

FIG. 6. Errors in interaction energies of chlorinated (top) or methylated (middle) ligands and a dipeptide modeled with different charge schemes. Errors are relative to fully QM interaction energies. The capped dipeptide contains the GLN and GLU residues, and GLU is modeled with point charges. The x-axis is the minimum distance between the Me ligand and the closest point charge (via BRC). The gray bar represents ± 1 kJ mol^{-1}. The bottom panel shows the difference in interaction energies, ΔE_{int}(Cl) – ΔE_{int}(Me).

kJ mol^{-1} with BRC2. DZ2, DZ3, BRC, BRC2, and BRCD give interaction energies within 1 kJ mol^{-1} for distances 7.7 Å and above. Z1 and Z2 perform especially poorly, and Z1 gives larger errors than even no embedding.

The second panel of Figure 6 is very similar, except the chlorine of the ligand is replaced with a methyl group. The dipeptide is identical to the earlier case, and now the intermolecular distance refers to the distance between the carbon of this methyl group and the closest point charge. Here, trends are similar, but the errors are larger in magnitude relative to the chlorinated ligand. This could be partially due to the larger QM interaction energy. At 6.7 Å, the interaction energy is –12.88 kJ mol^{-1}. BRC performs the best while small errors also occur for BRC2, BRCD, and DZ3. The worst performances are from DZ1, Z1, Z2, and no embedding.

Figure 6 also reports the error in the relative interaction energies of these two systems, subtracting the interaction energy of the methyl complex from that of the chlorine complex. Comparing energies of systems differing by a ligand substitution is a practice common in structure-based drug design.[43] BRC, DZ3, and BRCD produce relative interaction energy errors within 1 kJ mol^{-1} for all distances, and errors decrease at larger distances. Still, Z1, Z2, and no embedding are unsuitable models.

The same study has been conducted for dipeptides composed of the other residues mentioned. The systems are chosen to be GLU(HIE) (–1,0), HIP(HIE) (+1,0), GLN(HIE) (0,0), and GLN(HIP) (0,+1), where the first residue listed is closest to the ligand and modeled with QM and the second listed is modeled with MM. Respective charges follow the three-letter codes in parentheses. Corresponding graphs for these systems are available in the Supplementary Material, as well as a table of MM region charges and interaction energies for each system and charge scheme. An overview of the performance of different schemes is shown in Figure 7, where for each system and embedding scheme, we report the error in interaction energies averaged over the intermolecular separation. Overall, Z1 and Z2 produce the largest errors. Z1 overbinds due to the deletion of the M1 charge which makes the MM region more negative, increasing the strength of the interaction (making the interaction energy more negative) especially between the dipeptides and the methylated ligand. Alternatively, Z2 causes the opposite effect due to the deletion of M1 and M2 charges making the MM region more positive. Z3 returns the MM residue closer to its original charge, and in fact, Z3 yields lower errors than Z2 or Z1 in all but one case, GLN(HIP). In GLN(HIP), Z2 and Z3 perform similarly and the charges of their MM regions only differ by 0.01 a.u. (see Table S-8 for charges). These results support the conclusion that QM/MM charge assignment schemes that fail to conserve the charge in the MM region can lead to large errors.

All other charge schemes considered here maintain the original residue charge. DZ3 has lower average errors than DZ2, which has lower errors than DZ1. In each of these schemes, the charges needed to return the residue charge to its original integer are evenly distributed to the rest of the MM residue. The decrease in errors from DZ1 to DZ3 correlates with the amount of charge being distributed (see Table S-9).

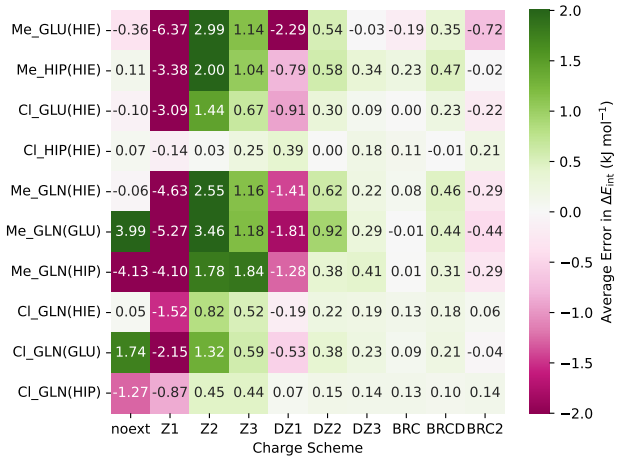

FIG. 7. Average of errors in embedded systems, relative to fully QM systems, with ligand to point charge closest contact distances of of 5.7 Å to 19.7 Å in 1 Å increments. Ligand substituents (Cl or Me) are noted in x-axis labels, and residues represented with point charges are in parentheses. The label 'noext' refers to no embedding (i.e. the residue in parentheses is not present). Interaction energies are calculated with FI-SAPT0/jun-cc-pV(D+d)Z, and the embedding scheme is noted on the x-axis.

The less the distributed charge, the lower the errors. As can be seen in GLN(HIP), when the distributed charge is about equal, the errors are about equal. For the BRC methods (BRC, BRCD, and BRC2), integer charge is maintained and the distributed charge is equal to that of DZ1. The difference here is how the charge is distributed. BRC usually performs the best, and it almost always outperforms DZ3 even though the distributed charge is larger. This is perhaps because BRC does not change the remaining point charges of the residue, but distributes the balancing charge to M1–M2 bond midpoints, and therefore causes very small changes in the charge distribution of the MM region.

When moving from these test systems to a more realistic protein, the increasing number of protein atoms available to interact with the protein at a given distance may require many cut bonds depending on the selection of the QM region. Thus, the smaller the error per frontier bond, the better, as errors could quickly accumulate. Overall, these results point to BRC producing the most accurate interaction energies for a small dipeptide and a ligand. Other methods expected to perform well are DZ3, BRCD, and BRC2, while Z1, Z2, Z3, and DZ1 are expected to perform poorly.

D. Protein-ligand interactions with EE-SAPT: Application to 2CJI

The protein-ligand system chosen as a test system, PDB ID 2CJI,[28] is prepared with the Protein Preparation Wizard[53,54] of Schrödinger's Maestro,[55] and details regarding this procedure can be found in the electronic Supplementary Material. For this complex, we study the same pair of truncated ligands used above: the chlorinated ligand of the original crystal structure (Cl), and a ligand where this chlorine has been replaced with a methyl group (Me). Our group's previous F-SAPT study of this complex truncated the protein to 142 atoms.[43] Now with EE-SAPT, we are able to uncover the effects of the rest of the protein ($\sim$ 4000 atoms) on the interaction energy.

For the 2CJI complex, we compute the interaction energy using each of the charge schemes and an increasingly larger QM region of the protein. To create each QM region, residues

Region	N_{resi}	N_{HL}	$N_{\text{QM at.}}$	q_{QM}	$d_{\text{lig.}-\text{MM}}$
1	3	4	46	0	2.4
2	7	8	97	0	2.5
3	10	10	142	-1	2.7
4	13	8	190	-2	2.7
5	14	8	214	-2	4.2
6	19	12	273	-2	4.6
7	24	14	352	-4	5.1
8	27	18	392	-4	5.5
9	30	20	446	-3	5.6
10*	32	23	490	-2	5.6
11*	36	26	551	-2	5.6
12*	39	26	588	-2	5.6

TABLE I. A description of the different chosen QM regions of the protein, including the number of residues (N_{resi}), hydrogen link atoms (N_{HL}), and QM atoms ($N_{\text{QM at}}$). The charge of the QM region (q_{QM}) and the distance from the methylated ligand to the closest MM point charge ($d_{\text{lig.}-\text{MM}}$) are included. $d_{\text{lig.}-\text{MM}}$ depends on the charge scheme; values are provided for BRC, and are thus minima compared to other charge schemes. Asterisks denote QM regions for which only electrostatics was computed. Images and Cartesian coordinates of these systems are available in the Supplementary Material.

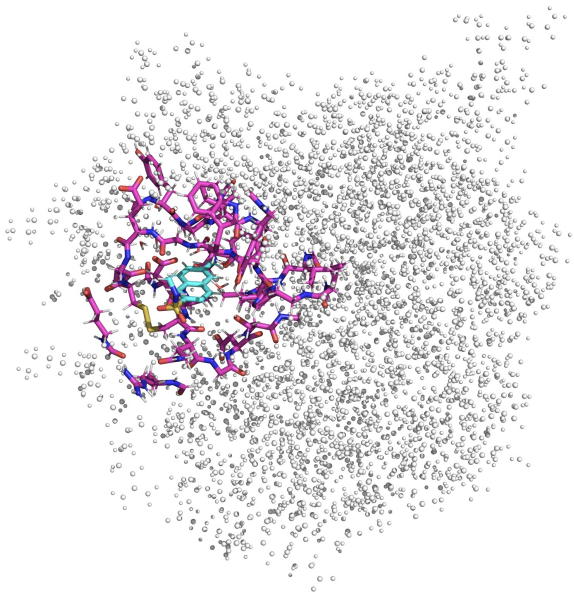

FIG. 8. 2CJI with QM region 9 in pink, the ligand in blue, and the rest of the protein represented by gray dots that signify point charges.

with at least one sidechain atom within a defined distance of the ligand fragment are selected. Then, the QM region changes until each frontier bond is a carbon-carbon bond, specifically the bond between a peptidic carbonyl carbon and the C_α, like in Fig. 5. The QM region may expand or shrink in order to cut the closest of these bonds. We have intentionally chosen to not cut polar bonds, and we do not cut peptide bonds due to their partial double-bond character. The QM regions of the protein range in size from 46 to 588 atoms, and details of these subsystems are listed in Table I. Full SAPT0 calculations on the systems larger than 446 atoms started to become very challenging computationally, so we turned to computing the electrostatic interaction energies only for regions 10–12. Illustrations of all regions are available in the Supplementary Material, and Figure 8 shows the largest region for which we were able to compute all four SAPT energy components.

In every calculation, except those labeled 'noext', the protein atoms not included in the QM region are included in the MM region (other than those explicitly zeroed for the charge scheme) such that the whole protein is accounted for, and this is shown in Figure 8. Water molecules with a minimum of one hydrogen bond to non-water molecules have been included by the protein preparation, and they are considered part of the protein (i.e. part of monomer B and modeled by QM if within the distance cutoff). Point charges are obtained through AmberTools22[56] using ff19SB[6] for the protein and OPC[57] for water molecules. Each water molecule has the same set of atomic charges.

Table II lists the interaction energies of each system when 446 protein atoms are included in the QM region, the system represented in Figure 8. The charge schemes return interaction energies between –60.40 and –55.44 kcal mol^{-1} for the chlorinated system, and between –60.16 and –52.04 kcal mol^{-1} for the methylated system. These ranges are large, about 5 and 8 kcal mol^{-1}, and for each system, and the lowest interaction energy is produced by Z1 and the highest is produced by Z2. From the dipeptide system, we know that significant errors can be present in interaction energies when using elimination schemes, no matter the distances between the two monomers. Ignoring the elimination schemes, the ranges of interaction energies using only the balanced schemes decrease to 2.5 and 2.6 kcal mol^{-1} for Cl and Me, respectively. Still, the balanced schemes do not totally agree to within chemical accuracy (1 kcal mol^{-1}) for these systems. For the relative interaction energies, the balanced schemes predict between –1.02 and –1.96 kcal mol^{-1}, a range of 0.9 kcal mol^{-1}.

Unable to compute a SAPT0 interaction energy for the whole protein, we turn to analyzing convergence in order to choose a recommended charge scheme and QM region size. For the Cl and Me systems, we define convergence as an energy within 1 kcal mol^{-1} of the best available value (from the largest QM region) for each given charge scheme. The relative energies are an order of magnitude smaller, so we will consider a converged $\Delta\Delta E$(Cl–Me) to be within 1 kJ mol^{-1} (0.24 kcal mol^{-1}) of the best available value. In the following, we analyze convergence on a per-energy-component basis. We expect the charge schemes to show more variance for electrostatics and induction than for dispersion or exchange, as electrostatics and induction are directly affected by external charges.

	Cl/Total	Me/Total	Cl–Me/Total
Z1	–60.40	–60.16	–0.24
Z2	–55.44	–52.04	–3.40
Z3	–57.69	–55.34	–2.35
DZ1	–56.90	–55.88	–1.02
DZ2	–57.55	–55.60	–1.96
DZ3	–58.56	–56.81	–1.75
BRC	–56.60	–54.95	–1.65
BRC2	–57.12	–55.62	–1.50
BRCD	–56.06	–54.25	–1.81
noext	–58.42	–55.72	–2.71

TABLE II. Interaction energies (kcal mol^{-1}) of the 2CJI protein and chlorinated (Cl) and methylated (Me) ligands computed with SAPT0/jun-cc-pV(D+d)Z and electrostatic embedding. The final column lists their relative interaction energies, ΔE_{int}(Cl) – ΔE_{int}(Me). 446 atoms of the protein are modeled by QM, and the rest of the protein is modeled with point charges. Charges at the boundary are handled differently, indicated by the charge scheme in each row. The final row, 'noext', presents interaction energies without external charges (QM regions only).

Figure 9 shows the convergence of each energy component for both the Cl and Me systems when using the different charge schemes as system size increases. It also shows the convergence of the relative energies. Importantly, the schemes do not always converge to the same point, and a table of energies for each scheme, component, and QM system size is provided in the Supplementary Material. The largest disagreement for the largest QM region is for electrostatics, where the balanced schemes compute ΔE_{elst}(Cl) to be between –33.15 and –31.34 kcal mol^{-1}, a range of 1.81 kcal mol^{-1}. For this chlorinated ligand, other ranges are: 0.10 (exchange), 0.85 (induction), and 0.07 (dispersion) kcal mol^{-1}. These range values vary by less than 0.1 kcal mol^{-1}, compared to Cl, when analyzing the methylated ligand.

Electrostatics, displayed in the upper left quadrant of Figure 9, certainly benefits from charge embedding, yet it is the most dependent on which scheme is chosen. The elimination schemes do not converge to within 1 kcal mol^{-1} for Me, and they barely converge for Cl. On the other hand, all of the balanced schemes do converge to within 1 kcal mol^{-1}, but at varying rates. For both cases, the fastest to converge within chemical accuracy is BRC. It computes the electrostatics interaction energies within 1 kcal mol^{-1} of the 558-atom systems by 190 atoms. The relative interaction energy, $\Delta\Delta E_{\text{elst}}$(Cl–Me), converges to within 1 kJ mol^{-1} of a scheme's electrostatic energy at 588 atoms when the balanced methods are used. Of the balanced schemes, DZ2 and DZ3 are the slowest to converge to 1 kJ mol^{-1}, needing the QM region to have 446 protein atoms. DZ1 converges a bit faster, but BRC, BRC2, and BRCD all converge to 1 kJ mol^{-1} more quickly than any of the the DZ schemes. BRC converges the fastest, beginning at 190 QM protein atoms. Mentioned previously, the low computational cost of electrostatics allows for electrostatics-only SAPT calculations on larger systems. Even

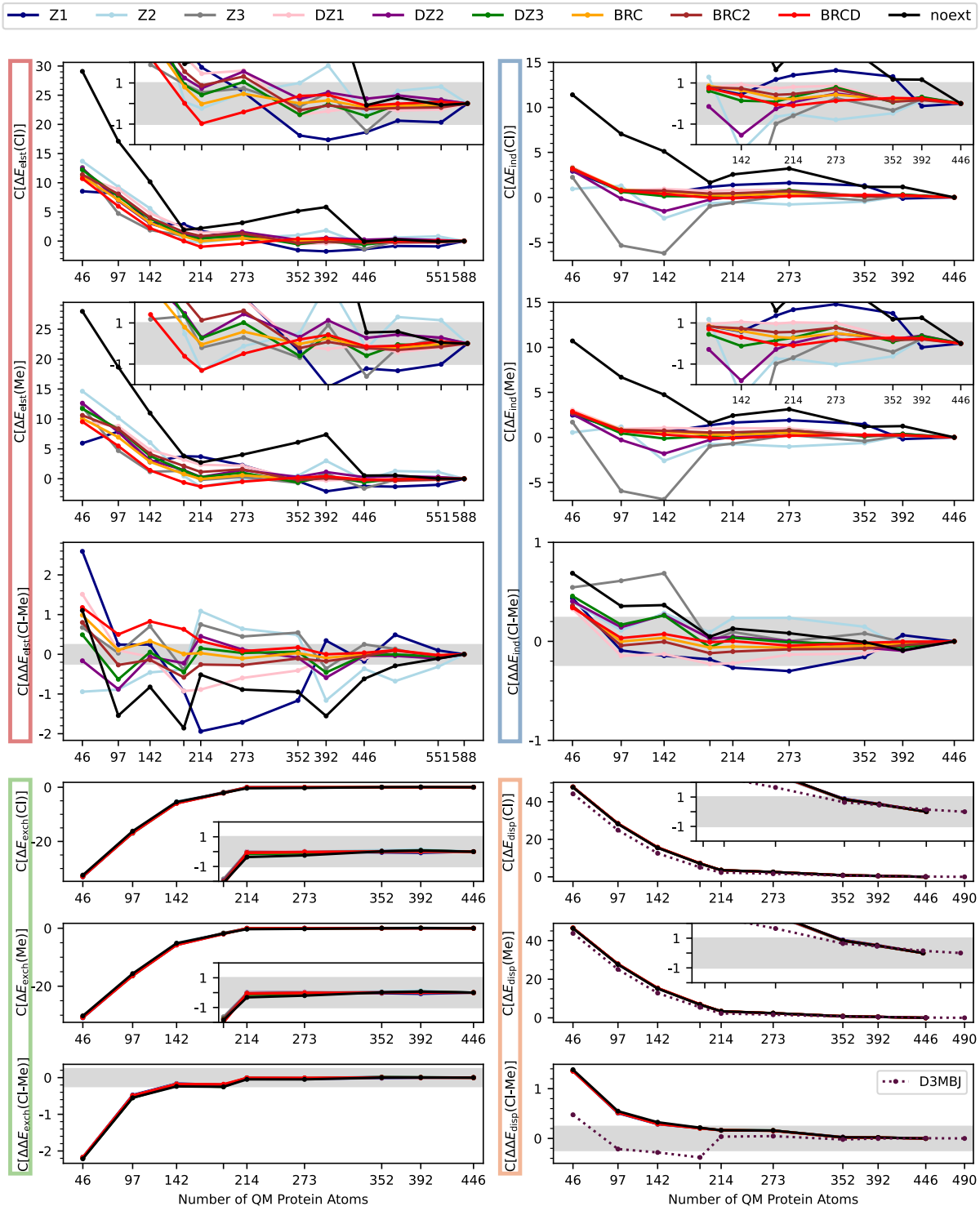

FIG. 9. Convergence (C) in kcal mol^{-1} of electrostatically embedded SAPT0/jun-cc-pV(D+d)Z components for 2CJI protein-ligand systems with chlorinated (Cl) and methylated (Me) ligands. The four quadrants represent electrostatics (upper left), induction (upper right), exchange (lower left), and dispersion (lower right). All charge schemes are represented, and convergence is evaluated in reference to the largest QM region for each scheme. The bottom graph of each quadrant shows convergence of the relative energies. In graphs of C[ΔE(Cl)] and C[ΔE(Me)], gray bars represent $\pm$ 1 kcal mol^{-1}. In graphs of C[$\Delta\Delta E$(Cl–Me)], gray bars represent $\pm$ 1 kJ mol^{-1}.

though we computed electrostatics for three systems larger than 446 atoms, we have seen that BRC converges much earlier.

Induction, shown in the upper right quadrant of Figure 9, can also be scheme-dependent, though slightly less so than electrostatics. Again, embedding accelerates convergence, especially when balanced schemes are used. BRC, BRC2, BRCD, and DZ3 compute induction energies within 1 kcal mol^{-1} of their respective reference values (with 446 atoms) by a QM protein region as small as 97 atoms for both systems. Relative induction energies, $\Delta\Delta E_{\text{ind}}$(Cl–Me) are converged to within 1 kJ mol^{-1} by 190 atoms when using any scheme except Z1. Convergence can be accelerated by using DZ1, BRC, BRCD, or BRC2, which all converge to within 1 kJ mol^{-1} by 97 atoms.

Both electrostatics and induction show a large dependence on QM region size when a system is truncated and does not include the rest of the protein via point charges ('noext'). With no embedding, electrostatics does not converge to within 1 kcal mol^{-1} of its 588 atom limit for either ligand until 446 atoms are included (recall that, for electrostatics only, we were able to evaluate QM values for up to 588 atoms). For induction, convergence within 1 kcal mol^{-1} of its 446-atom reference value was not reached by the next-largest QM model system, 392 atoms. These results suggest that calculating a well-converged interaction energy without embedding requires at least 446 atoms in the protein's QM region. Even for the relative interaction energy, electrostatics is slow to converge without electrostatic embedding and requires 551 protein atoms for a $\Delta\Delta E_{\text{elst}}$ within 1 kJ mol^{-1} of the electrostatics energy of the 588 atom system.

The bottom half of Figure 9 is much more simple than the top half. For exchange and dispersion, all schemes generally agree, and they produce results very similar to those without embedding. Whereas electrostatics and induction are directly affected by external charges, exchange and dispersion are only indirectly affected, accounting for the change in monomer density due to the embedding charges. In both the chlorine and methyl systems, exchange is converged to within 1 kcal mol^{-1} of the system with the largest QM region by 214 protein QM atoms, and the relative energy is converged to 1 kJ mol^{-1} by 142 atoms (except for DZ3 which is just outside 1 kJ mol^{-1} until 214 QM protein atoms are included). Dispersion takes the longest to converge, waiting until 352 atoms to be within 1 kcal mol^{-1} of the dispersion energy of the 446-atom systems for Me and Cl. We have also computed the dispersion energy of each QM region with SAPT0-D3, which replaces the dispersion terms with reparametrized semi-empirical -D3 terms (-D3MBJ).[18] The convergence of D3MBJ is not notably faster than that of SAPT0's dispersion, but it does suggest that the dispersion energy barely changes when expanding the QM region to 490 atoms. $\Delta\Delta E_{\text{disp}}$(Cl–Me) converges to within 1 kJ mol^{-1} of the largest QM region by 190 QM protein atoms.

Due to the excellent convergence of BRC in all four components and its performance for the dipeptide systems, we recommend its use when preparing a complex for SAPT0 calculations with electrostatic embedding. Figure 10 shows the component and total interaction energies for both the chlorine and methyl systems, as well as the differences in energies, when using BRC. At the largest QM region of 446 atoms, ΔE_{int}(Cl) = –56.60 kcal mol^{-1} and ΔE_{int}(Me) = –54.95 kcal mol^{-1}. The chlorine system is more stable than the methyl system by –1.65 kcal mol^{-1} which happens to be in excellent agreement with the experimental $\Delta\Delta G_{\text{bind}}$ = –1.7 kcal mol^{-1}.[28] Note that our calculation is lacking some components of $\Delta\Delta G_{\text{bind}}$, like solvation and deformation energies. Because we have only made a minor change to the ligand substituent, we expect these additional terms to largely cancel when evaluating the relative energies.[43]

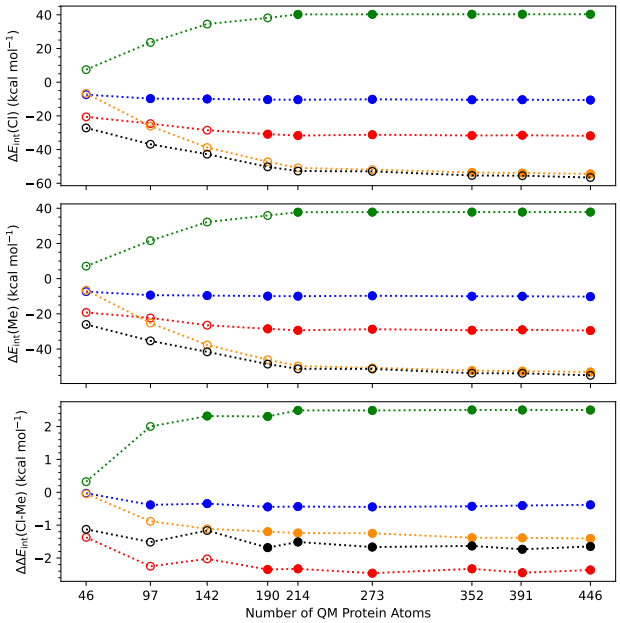

FIG. 10. Interaction energies (black) for 2CJI and the chlorinated (Cl) and methylated (Me) ligands, and their relative interaction energies. Protein atoms not included in the QM region are represented by point charges, and the boundary region is altered according to the balanced redistributed charge (BRC) scheme. Components are also presented: exchange (green), induction (blue), electrostatics (red), and dispersion (orange). Markers are filled in once the value is within 1 kcal mol^{-1} of the final value for ΔE(Cl) and ΔE(Me), and once the value is within 1 kJ mol^{-1} of the final value for $\Delta\Delta E$(Cl–Me).

Figure 10 also indicates, via a filled in marker, how large the QM region needs to be to reach convergence within 1 kcal mol^{-1} of the interaction energy of the largest QM region for each component, and within 1 kJ mol^{-1} for each relative interaction energy. For all four components to be simultaneously converged to within 1 kcal mol^{-1} for the Me and Cl complexes, 352 QM atoms are needed in the protein for 2CJI. This region has 24 residues, and the closest point charge to the ligand is 5.1 Å away (see Table I). While the total interaction energy itself does not yet converge to within chemical accuracy, for this 352 QM atom system, it is within 1.22 (Me) or 1.16 (Cl) kcal mol^{-1} of the final interaction energy (446 atoms). The relative interaction energies require that only 190 protein atoms be included in the QM regions for convergence within 1 kJ mol^{-1} for all four components and the total inter-

action energy.

IV. CONCLUSIONS

In this article, we reported the extension of electrostatic embedding to symmetry-adapted perturbation theory. We illustrated possible embedding scenarios for our 3-subsystem scheme (A, B, C) using water hexamer and showed that partial charges can be added to the external environment (C) or to the interacting subsystems (A and B). We analyzed the various embedding schemes using water trimer as a prototype example and demonstrated the long-range convergence of the embedding procedure.

We have also shown that SAPT with electrostatic embedding works well for practical use cases. The method can calculate the interaction energy of the solvated pyridinium:benzene and benzene dimers with an accuracy of 1 kJ mol^{-1} when only the closest 18 waters are modeled with QM and the remaining waters are modeled with MM. EE-(I)SAPT does suffer from the general limitations of electrostatic embedding, such as the neglect of mutual polarization and the overpolarization of the QM region by the MM potential. We test embedded SAPT on systems where the QM/MM boundary spans covalent bonds and have shown that non-neglible errors may be introduced by the cut-and-cap procedure. For model dipeptide-ligand systems, the charge schemes DZ3, BRC, BRCD, and BRC2 returned lower interaction energy errors relative to other schemes tested, and we believe embedded SAPT will benefit from future advances in QM/MM partitioning.[1,2]

SAPT0 can now be applied to larger systems with thousands of atoms, and we computed interaction energies of the factor Xa protein (PDB 2CJI) with two similar ligands. Relative to no embedding, SAPT with electrostatic embedding increased the speed of convergence of the interaction energy (in particular, the electrostatics and induction components) as QM system size is increased. These two components converged the fastest when using BRC rather than the other tested charge schemes. Exchange and dispersion showed similar convergence behavior with all of the charge schemes. Using electrostatic embedding, the BRC scheme, and a QM region with 352 protein atoms, ΔE_{int}(Cl) = –56.60 kcal mol^{-1} and ΔE_{int}(Me) = –54.95 kcal mol^{-1} for 2CJI. Often the relative binding of two ligands is of the most interest, and in this case the relative interaction energy, $\Delta\Delta E_{\text{int}} = -1.65$ kcal mol^{-1}, is very close to the experimental $\Delta\Delta G_{\text{bind}}$ of –1.7 kcal mol^{-1}. This close agreement is likely due to an expected near cancellation of differential ligand solvation energies and entropies. A substantially less expensive computation involving only 190 QM protein atoms (with the rest of the protein again modeled by point charges) yields an essentially identical $\Delta\Delta E_{\text{int}}$ = –1.68 kcal mol^{-1}. These results, and others presented in the paper, suggest that electrostatic embedding is an effective way to greatly speed up SAPT analysis by replacing distant atoms with point charge representations, with minimal impact on accuracy. EE-(I)SAPT will make SAPT analysis of protein-ligand interactions, solvated systems, or other complex systems much more tractable.

SUPPLEMENTARY MATERIAL

See the supplementary material for detailed SAPT0 interaction energy tables for water trimer and protein-ligand systems, 2D structures of the ligand, and illustrations of the 2CJI QM regions. The electronic Supplementary Material provides Psi4 input and output files for the solvated dimer systems. It also contains PDBs for the protein-ligand systems, mol2 files, and csv files with interaction energy results, and details on the protein preparation steps for the protein-ligand tests.

ACKNOWLEDGMENTS

This research was supported by U.S. National Science Foundation grants to C.D.S. (CHE-1955940 and ACI-1449723) and a NSF Graduate Research Fellowship to C.S.G. (DGE-2039655). C.C. was supported by the National Institutes of Health (NIH), through a T2 training grant (32GM135134) to Lynne Maquat and Jeff Hayes, and through an R01 grant (R01GM132185) to David Mathews. Some of the computations were performed on the Georgia Tech Hive Cluster, funded through NSF grant MRI-1828187, and maintained by the Partnership for an Advanced Computing Environment (PACE) at Georgia Tech.

AUTHOR DECLARATIONS

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

[1] Q. Cui, T. Pal, and L. Xie, "Biomolecular QM/MM simulations: What are some of the burning issues?" J. Phys. Chem. B **125**, 689–702 (2021).
[2] M. Bondanza, M. Nottoli, L. Cupellini, F. Lipparini, and B. Mennucci, "Polarizable embedding QM/MM: the future gold standard for complex (bio)systems?" Phys. Chem. Chem. Phys. **22**, 14433–14448 (2020).
[3] A. O. Dohn, "Multiscale electrostatic embedding simulations for modeling structure and dynamics of molecules in solution: A tutorial review," Int. J. Quantum Chem. **120**, e26343 (2020).
[4] J. Huang and A. D. MacKerell, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data," J. Comput. Chem. **34**, 2135–2145 (2013).
[5] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "Ff14SB: Improving the accuracy of protein side chain and backbone parameters from Ff99SB," J. Chem. Theory Comput. **11**, 3696–3713 (2015).
[6] C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Z. Wang, J. Pincay, Q. Wu, and C. Simmerling, "ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution," J. Chem. Theory Comput. **16**, 528–552 (2020).

[7]B. Jeziorski, R. Moszynski, and K. Szalewicz, "Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes," Chem. Rev. **94**, 1887–1930 (1994).

[8]K. Patkowski, "Recent developments in symmetry-adapted perturbation theory," WIRES Comput Mol Sci. **10**, e1452 (2020).

[9]R. M. Parrish, J. F. Gonthier, C. Corminboeuf, and C. D. Sherrill, "Communication: Practical intramolecular symmetry adapted perturbation theory via hartree-fock embedding," J. Chem. Phys. **143**, 051103 (2015).

[10]R. M. Parrish and C. D. Sherrill, "Spatial assignment of symmetry adapted perturbation theory interaction energy components: The atomic SAPT partition," J. Chem. Phys. **141**, 044115 (2014).

[11]R. M. Parrish, T. M. Parker, and C. D. Sherrill, "Chemical assignment of symmetry-adapted perturbation theory interaction energy components: The functional-group SAPT partition," J. Chem. Theory Comput. **10**, 4417–4431 (2014).

[12]E. G. Hohenstein and C. D. Sherrill, "Density fitting and Cholesky decomposition approximations in symmetry-adapted perturbation theory: Implementation and application to probe the nature of π-π interactions in linear acenes," J. Chem. Phys. **132**, 184111 (2010).

[13]F. Rob, R. Podeszwa, and K. Szalewicz, "Electrostatic interaction energies with overlap effects from a localized approach," Chem. Phys. Lett. **445**, 315–320 (2007).

[14]F. Rob, A. J. Misquitta, R. Podeszwa, and K. Szalewicz, "Localized overlap algorithm for unexpanded dispersion energies," J. Chem. Phys. **140**, 114304 (2014).

[15]K. U. Lao and J. M. Herbert, "Accurate and efficient quantum chemistry calculations for noncovalent interactions in many-body systems: The XSAPT family of methods," J. Phys. Chem. A **119**, 235–252 (2015).

[16]A. Hesselmann, "Comparison of intermolecular interaction energies from SAPT and DFT including empirical dispersion contributions," J. Phys. Chem. A **115**, 11321–11330 (2011).

[17]R. Sedlak and J. Řezáč, "Empirical D3 dispersion as a replacement for ab lnitio dispersion terms in density functional theory-based symmetry-adapted perturbation theory," J. Chem. Theory Comput. **13**, 1638–1646 (2017).

[18]J. B. Schriber, D. A. Sirianni, D. G. A. Smith, L. A. Burns, D. Sitkoff, D. L. Cheney, and C. D. Sherrill, "Optimized damping parameters for empirical dispersion corrections to symmetry-adapted perturbation theory," J. Chem. Phys. **154**, 234107 (2021).

[19]K. U. Lao and J. M. Herbert, "A simple correction for nonadditive dispersion within extended symmetry-adapted perturbation theory (XSAPT)," J. Chem. Theory Comput. **14**, 5128–5142 (2018).

[20]K. U. Lao and J. M. Herbert, "Atomic orbital implementation of extended symmetry-adapted perturbation theory (XSAPT) and benchmark calculations for large supramolecular complexes," J. Chem. Theory Comput. **14**, 2955–2978 (2018).

[21]A. Heßelmann, G. Jansen, and M. Schütz, "Density-functional theory-symmetry-adapted intermolecular perturbation theory with density fitting: A new efficient method to study intermolecular interaction energies," J. Chem. Phys. **122**, 014103 (2005).

[22]A. J. Misquitta, R. Podeszwa, B. Jeziorski, and K. Szalewicz, "Intermolecular potentials based on symmetry-adapted perturbation theory with dispersion energies from time-dependent density-functional calculations," J. Chem. Phys. **123**, 214103 (2005).

[23]J. G. McDaniel and J. R. Schmidt, "Next-generation force fields from symmetry-adapted perturbation theory," Annu. Rev. Phys. Chem. **67**, 467–488 (2016).

[24]J. G. McDaniel and J. R. Schmidt, "Physically-motivated force fields from symmetry-adapted perturbation theory," J. Phys. Chem. A **117**, 2053–2066 (2013).

[25]D. P. Metcalf, A. Koutsoukas, S. A. Spronk, B. L. Claus, D. A. Loughney, S. R. Johnson, D. L. Cheney, and C. D. Sherrill, "Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory," J. Chem. Phys. **152**, 074103 (2020).

[26]Z. L. Glick, D. P. Metcalf, A. Koutsoukas, S. A. Spronk, D. L. Cheney, and C. D. Sherrill, "AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials," J. Chem. Phys. **153**, 044112 (2020).

[27]J. B. Schriber, D. R. Nascimento, A. Koutsoukas, S. A. Spronk, D. L. Cheney, and C. D. Sherrill, "CLIFF: A component-based, machine-learned, intermolecular force field," J. Chem. Phys. **154**, 184110 (2021).

[28]C. Chan, A. D. Borthwick, D. Brown, C. L. Burns-Kurtis, M. Campbell, L. Chaudry, C. Chung, M. A. Convery, J. N. Hamblin, L. Johnstone, H. A. Kelly, S. Kleanthous, A. Patikis, C. Patel, A. J. Pateman, S. Senger, G. P. Shah, J. R. Toomey, N. S. Watson, H. E. Weston, C. Whitworth, R. J. Young, and P. Zhou, "Factor Xa inhibitors: S1 binding interactions of a series of n-(3S)-1-[(1S)-1-methyl-2-morpholin-4-yl-2-oxoethyl]-2- oxopyrrolidin-3- Ylsulfonamides," J. Med. Chem. **50**, 1546–1557 (2007).

[29]T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno, and C. D. Sherrill, "Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies," J. Chem. Phys. **140**, 094106 (2014).

[30]S. F. Boys and F. Bernardi, "The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors," Mol. Phys. **19**, 553–566 (1970).

[31]G. Knizia, "Intrinsic atomic orbitals: An unbiased bridge between quantum theory and chemical concepts," J. Chem. Theory Comput. **9**, 4834–4843 (2013).

[32]C. Perez, M. T. Muckle, D. P. Zaleski, N. A. Seifert, B. Temelso, G. C. Shields, Z. Kisiel, and B. H. Pate, "Structures of cage, prism, and book isomers of water hexamer from broadband rotational spectroscopy," Science **336**, 897–901 (2012).

[33]D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.4: Open-source software for high-throughput quantum chemistry," J. Chem. Phys. **152**, 184108 (2020).

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

[35]C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model," J. Phys. Chem. **97**, 10269–10280 (1993).

[36]A. Alenaizan, L. A. Burns, and C. D. Sherrill, "Python implementation of the restrained electrostatic potential charge model," Int. J. Quantum Chem. **120**, e26035 (2020).

[37]T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. V. Speybroeck, M. Waroquier, and P. W. Ayers, "Minimal basis iterative Stockholder: Atoms in molecules for force-field development," J. Chem. Theory Comput. **12**, 3894–3912 (2016).

[38]J. Segarra-Martí, M. Merchán, and D. Roca-Sanjuán, "Ab initio determination of the ionization potentials of water clusters $(H_2O)_n$ (n=2-6)," J. Chem. Phys. **136**, 244306 (2012).

[39]T. M. Parker and C. D. Sherrill, "Assessment of empirical models versus high-acuracy ab initio methods for nucleobase stacking: Evaluating the importance of charge penetration," J. Chem. Theory Comput. **11**, 4197–4202 (2015).

[40]D. A. Sirianni, X. Zhu, D. F. Sitkoff, D. L. Cheney, and C. D. Sherrill, "The influence of a solvent environment on direct non-covalent interactions between two molecules: A symmetry-adapted perturbation theory study of polarization tuning of π-π interactions by water," J. Chem. Phys. **156**, 194306 (2022).

[41]J. Řezáč, K. E. Riley, and P. Hobza, "S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures," J. Chem. Theory Comput. **7**, 2427–2438 (2011).

[42]J. Řezáč, K. E. Riley, and P. Hobza, "Erratum to S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures," J. Chem. Theory Comput. **10**, 1359–1360 (2014).

[43]R. M. Parrish, D. F. Sitkoff, D. L. Cheney, and C. D. Sherrill, "The surprising importance of peptide bond contacts in drug-protein interactions," Chem. Eur. J. **23**, 7887–7890 (2017).

[44]G. E. Merz, M. J. Chalkley, S. K. Tan, E. Tse, J. N. Lee, S. B. Prusiner, N. A. Paras, W. F. DeGrado, and D. R. Southworth, "Stacked binding of a PET ligand to alzheimer's tau paired helical filaments," Nat. Commun. **14**, 3048 (2023).

[45]H. Lin and D. G. Truhlar, "Qm/mm: What have we learned, where are we, and where do we go from here?" Theor. Chem. Acc. **117**, 185–199 (2007).
[46]H. M. Senn and W. Thiel, "QM/MM methods for biomolecular systems," Angew. Chem. Int. Ed. **48**, 1198–1229 (2009).
[47]L. H. Hu, P. Söderhjelm, and U. Ryde, "On the convergence of QM/MM energies," J. Chem. Theory Comput. **7**, 761–777 (2011).
[48]H. Lin and D. G. Truhlar, "Redistributed charge and dipole schemes for combined quantum mechanical and molecular mechanical calculations," J. Phys. Chem. A **109**, 3991–4004 (2005).
[49]M. Isegawa, B. Wang, and D. G. Truhlar, "Electrostatically embedded molecular tailoring approach and validation for peptides," J. Chem. Theory Comput. **9**, 1381–1393 (2013).
[50]B. Wang and D. G. Truhlar, "Combined quantum mechanical and molecular mechanical methods for calculating potential energy surfaces: Tuned and balanced redistributed-charge algorithm," J. Chem. Theory Comput. **6**, 359–369 (2010).
[51]B. Waszkowycz, I. H. Hillier, N. Gensmantel, and D. W. Payling, "A combined quantum mechanical/molecular mechanical model of the potential energy surface of ester hydrolysis by the enzyme phospholipase a2," J. Chem. Soc., Perkin Trans. 2 **-**, 225–231 (1991).
[52]U. C. Singh and P. A. Kollman, "A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl + Cl^-$ exchange reaction and gas phase protonation of polyethers," J. Comput. Chem. **7**, 718–730 (1986).
[53]G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, and W. Sherman, "Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments," J. Comput. Aid. Mol. Des. **27**, 221–234 (2013).
[54]Schrödinger Release 2023-3: Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2023; Impact, Schrödinger, LLC, New York, NY; Prime, Schrödinger, LLC, New York, NY, 2023.
[55]Schrödinger Release 2023-3: Maestro, Schrödinger, LLC, New York, NY, 2023.
[56]D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2022), Amber 2022, University of California, San Francisco.
[57]S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building water models: A different approach," J. Phys. Chem. Lett. **5**, 3863–3871 (2014).