Title
Advances in Density Functional Theory Calculations and Their Analysis

Author
Mao, Yuezhi

Publication Date
2017

Peer reviewed|Thesis/dissertation

ADVANCES IN DENSITY FUNCTIONAL THEORY CALCULATIONS AND THEIR ANALYSIS

by

YUEZHI MAO

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin P. Head-Gordon, Chair
Professor Eric W. Neuscamman
Professor Daryl C. Chrzan

Fall 2017

ADVANCES IN DENSITY FUNCTIONAL THEORY CALCULATIONS AND THEIR ANALYSIS

Abstract

ADVANCES IN DENSITY FUNCTIONAL THEORY CALCULATIONS AND THEIR ANALYSIS

by

YUEZHI MAO

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Martin P. Head-Gordon, Chair

This thesis is primarily concerned with the development of tools aiming to improve the efficiency of electronic structure calculations using modern density functional theory (DFT), and to utilize the results of DFT-based energy decomposition analysis (EDA) for probing and modeling intermolecular interactions. It covers three distinct while also related topics: (i) development of a novel scheme to perform self-consistent field (SCF) calculations using a minimal adaptive basis (MAB) that is optimized on-the-fly; (ii) development and assessment of energy decomposition analysis (EDA) methods, including a new EDA scheme that is capable of probing the observable effects of intermolecular interactions, and a thorough discussion on the definition of the charge-transfer term in EDA methods; (iii) application of EDA to the development of potential energy surfaces for molecular simulations, including an assessment of the AMOEBA polarizable force field, and the formulation, implementation, and analysis of a mutually polarizable QM/MM scheme employing AMOEBA.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank Prof. Martin Head-Gordon for being an incredibly helpful and supportive advisor. From him I learned a lot about electronic structure theory and ways to think about and address scientific problems in general. I am grateful to him for his invaluable advice on the research presented here, and for his encouragement and support during the 5+ years that made these accomplishments possible. They are definitely unforgettable experiences to discuss research with him in our weekly meetings and to become his TA for two graduate-level courses about quantum mechanics and computational chemistry.

I would like to gratefully acknowledge Profs. Teresa Head-Gordon, Chris-Kriton Skylaris, Yihan Shao, and Drs. Omar Demerdash, Jacek Dziedzic for being amazing collaborators on the work presented in the last two chapters. From Teresa and Omar I learned many details about polarizable force fields and molecular simulations, and from the conversations with Chris and Jacek I gained many useful insights into linear-scaling DFT and development of polarizable QM/MM schemes. Yihan's expertise in QM/MM and Q-Chem code development is indispensable for the emergence of the Q-Chem/LibEFP code interface, and it is always enjoyable to chat about science and other things with him.

I would like to thank Eric Sundstrom, Kristi Closser, Julian Azar, Paul Horn, David Stück, Samuel Manzer, Narbe Mardirossian, Jonathan Thirman, Nicholas Mayhall, and Andreas Hauser for offering me extensive helps and guides when I just joined the Head-Gordon group. In particular, I would like to thank Paul and Narbe whose help and scientific contributions have made this thesis work possible. Paul taught me the math tools that greatly facilitate the derivation of equations in a nonorthogonal basis. His prominent work on the theory and code development for the second-generation ALMO-EDA lays the foundation for the work presented in Chapters 3–6. Narbe taught me all the details about running DFT calculations with Q-Chem (the best one can know) and is always ready to provide helpful information about the accuracy of density functionals based on his incredibly extensive benchmark data. The excellent performance of the density functionals developed by him also motivates or validates the work in this thesis. It was truly lucky having the chance to ever work with them. I also want to thank Alec White, Jon Witte, Ehud Tsivion, Daniel Levine, Shane Yost, Jérôme Gonthier, Eloy Ramos-Cordoba, Xintian Feng, Luke Bertals, Joonho Lee, Srimukh Prasad, and Matthias Loipersberger, who are currently in or recently left the MHG group, for having helpful discussions and creating a friendly and supportive group atmosphere.

I would like to thank Qinghui Ge for being an amazing lab mate, collaborator, and companion in life. I enjoyed and benefited a lot from the collaboration with her on ALMO-based excited state methods, and truly appreciate her help on the work presented in Chapter 4 and the on-going efficient implementation of MAB-SCF. Meanwhile, the experiences of having lunch/dinner together, hiking in Muir Woods and other places, and driving to Oregon to see the marvelous total solar eclipse with her have made my graduate school life much more colorful and enjoyable.

I would like to thank Leslie Silvers and Jessie Woodcock for taking care of the administration issues such as employment, travel reimbursement, etc. I also want to thank Drs. Evgeny Epifanovsky, Zhi-Qiang You, and Zhengting Gan for their assistance on the development and use of the Q-Chem software package.

At last, I would like to thank my parents for their consistent love and support. It was always relaxing and joyful being at home, and I wish that I could visit them more often in the future.

Chapter 1

Introduction

1.1 The Electronic Schrödinger Equation

1.1.1 The Born-Oppenheimer Approximation

The Schrödinger equation proposed in 1920s describes the motion of atomic nuclei and electrons in the microscopic world. Given a system comprising N nuclei and n electrons, its energy and details about the electronic structure can be obtained by solving the time-independent Schrödinger equation:

$$\hat{H}\left|\Psi\right\rangle = E\left|\Psi\right\rangle, \tag{1.1}$$

where $\hat{H}$ is the Hamiltonian operator for this system, and the system energy (E) and the wavefunction ($|\Psi\rangle$) are given as an eigenvalue and the corresponding eigenfunction of $\hat{H}$, respectively. In the real space, the Hamiltonian operator can be expressed in atomic units as

$$\begin{aligned}\hat{H} &= \hat{T}_{\mathrm{e}} + \hat{T}_{\mathrm{n}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{nn}} \qquad (1.2)\\ &= -\sum_{i=1}^{n}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{N}\frac{1}{2m_A}\nabla_A^2 - \sum_{i=1}^{n}\sum_{A=1}^{N}\frac{Z_A}{|\mathbf{r}_i-\mathbf{R}_A|}\\ &\quad + \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} + \sum_{A=1}^{N}\sum_{B>A}^{N}\frac{Z_A Z_B}{|\mathbf{R}_A-\mathbf{R}_B|}, \qquad (1.3)\end{aligned}$$

where $\hat{T}_{\mathrm{e}}$ ($\hat{T}_{\mathrm{n}}$) is the kinetic energy operator for electrons (nuclei), $\hat{V}_{\mathrm{ne}}$, $\hat{V}_{\mathrm{ee}}$ and $\hat{V}_{\mathrm{nn}}$ represent the nuclei-electron, electron-electron and nuclei-nuclei Coulomb interactions, $\mathbf{r}_i$ corresponds to the coordinates of the electrons, and m_A, Z_A and $\mathbf{R}_A$ stand for the masses, charges and coordinates of the nuclei. Correspondingly, the wavefunction can be denoted as $\Psi(\mathbf{r},\mathbf{R})$, where $\mathbf{r} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_n\}$ and $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots \mathbf{R}_N\}$ are collective coordinates of electrons

and nuclei, respectively. According to the basic postulates of quantum mechanics, the modulus of the wavefunction, $|\Psi(\mathbf{r},\mathbf{R})|^2$, gives the *probability density* of finding the electrons at $\mathbf{r}$ and simultaneously the nuclei at $\mathbf{R}$.

The motions of electrons and nuclei are coupled in the Schrödinger equation given by Eqs. (1.1)–(1.3), which will be unfeasible to tackle for most systems of interest due to the complexity. A reasonable way to simplify the Schrödinger equation is the Born-Oppenheimer (BO) approximation, which is rationalized by the fact that nuclei are thousands of times heavier than electrons and are moving much more slowly. Under the BO approximation, the electrons are moving in the Coulomb potential of fixed nuclei, and the wavefunction of the system can be factorized as

$$\Psi(\mathbf{r},\mathbf{R}) = \Psi_{\mathrm{e}}(\mathbf{r};\mathbf{R})\Psi_{\mathrm{n}}(\mathbf{R}), \tag{1.4}$$

where $\Psi_{\mathrm{e}}(\mathbf{r};\mathbf{R})$ is the solution to the *electronic* Schrödinger equation (Eq. (1.5)) under a fixed configuration of nuclei ($\{\mathbf{R}_1,\mathbf{R}_2,\ \ldots\mathbf{R}_N\}$):

$$\left(-\sum_{i=1}^{n}\frac{1}{2}\nabla_i^2 - \sum_{i=1}^{n}\sum_{A=1}^{N}\frac{Z_A}{|\mathbf{r}_i-\mathbf{R}_A|} + \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}\right)\Psi_{\mathrm{e}}(\mathbf{r};\mathbf{R}) = E_{\mathrm{e}}(\mathbf{R})\Psi_{\mathrm{e}}(\mathbf{r};\mathbf{R}), \tag{1.5}$$

which can also be referred to as an *electronic state*. To solve Eq. (1.5) is a central problem in electronic structure theory.

An important concept associated with the BO approximation is the *potential energy surface* (PES) on which the nuclei are moving. It is a function of nuclear coordinates ($\mathbf{R}$):

$$V(\mathbf{R}) = E_{\mathrm{e}}(\mathbf{R}) + V_{\mathrm{nn}}(\mathbf{R}), \tag{1.6}$$

where $E_{\mathrm{e}}(\mathbf{R})$ is obtained by solving the electronic Schrödinger equation at each given nuclear configuration. The concept of PES is widely used in the following chapters, and it can easily be derived by substituting Eq. (1.4) into the original Schrödinger equation and applying Eq. 1.5, i.e., the Schrödinger equation for the nuclei has the following form:

$$(\hat{T}_{\mathrm{n}} + \hat{V}(\mathbf{R}))\Phi_{\mathrm{n}}(\mathbf{R}) = W\Phi_{\mathrm{n}}(\mathbf{R}). \tag{1.7}$$

The BO approximation could break down when the energies of two electronic states are close to each other (one well-known example is at the crossing point of the ionic and covalent states upon the dissociation of LiF). Nevertheless, it provides a good approximation for problems within the scope of this thesis, and for brevity, we use $\Psi_{\mathrm{e}}(\mathbf{r})$ instead of $\Psi_{\mathrm{e}}(\mathbf{r};\mathbf{R})$ to refer to the electronic wavefunction in the following discussion.

1.1.2 Hartree-Fock Theory

The electronic Schrödinger equation given by Eq. (1.5) is a $3n$-dimensional partial differential equation, whose analytical solution usually cannot be obtained for systems containing more than one single electron. Therefore, further approximations are usually needed to get

a reasonable numerical solution to the electronic structure problem with affordable cost. Thanks to the variational principle, the upper bound to the true ground state electronic energy ($\mathcal{E}_0$) can be accessed by minimizing the energy of a trial wavefunction $|\tilde{\Psi}\rangle$:

$$\tilde{E}_0 = \min_{\mathbf{c}} \frac{\langle\tilde{\Psi}(\mathbf{c})|\hat{H}|\tilde{\Psi}(\mathbf{c})\rangle}{\langle\tilde{\Psi}(\mathbf{c})|\tilde{\Psi}(\mathbf{c})\rangle} \geq \mathcal{E}_0, \tag{1.8}$$

where $\mathbf{c}$ represents the set of parameters to be optimized in the trial wavefunction.

The simplest trial wavefunction for an n-electron system is probably the antisymmetric product of n one-particle wavefunctions, i.e., the Slater determinant:

$$\Phi_0(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_1(\mathbf{x}_2) & \dots & \chi_1(\mathbf{x}_n) \\ \chi_2(\mathbf{x}_1) & \chi_2(\mathbf{x}_2) & \dots & \chi_2(\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_n(\mathbf{x}_1) & \chi_n(\mathbf{x}_2) & \dots & \chi_n(\mathbf{x}_n) \end{vmatrix}, \tag{1.9}$$

which satisfies the requirement of Fermionic statistics that the wavefunction changes to an opposite sign upon a single permutation:

$$\Phi_0(..., \mathbf{x}_i, ..., \mathbf{x}_j, ...) = (-1)\Phi_0(..., \mathbf{x}_j, ..., \mathbf{x}_i, ...). \tag{1.10}$$

Each one-particle wavefunction in the Slater determinant is referred to as a spin orbital, which has a spatial component and a spin component:

$$\chi_i(\mathbf{x}_j) = \psi_i(\mathbf{r}_j)\sigma_i(\omega_j), \tag{1.11}$$

where ψ_i is the ith spatial orbital, $\mathbf{r}_j$ is the spatial coordinate of the jth electron, and ω_j is its "conceptual" spin coordinate ($i, j = 1, 2, ..., n$). As an electron is a spin-1/2 particle ($m_s = \pm 1/2$), the spin states ($\sigma = \alpha$, β) can be chosen to be orthonormal:

$$\int \mathrm{d}\omega_j \sigma'(\omega_j)\sigma(\omega_j) = \delta_{\sigma\sigma'}. \tag{1.12}$$

The Hartree-Fock (HF) approximation employs the normalized Slater determinant (Φ_0) as the trial wavefunction for the variational optimization represented by Eq. (1.8):

$$E_{\mathrm{HF}} = \min_{\{\chi\}} \langle\Phi_0|\hat{H}_{\mathrm{e}}|\Phi_0\rangle\,, \tag{1.13}$$

where $\hat{H}_{\mathrm{e}}$ is the electronic Hamiltonian (the operator on the LHS of Eq. (1.5)), which can be further partitioned into a one-electron part

$$\hat{O}_1 = -\sum_{i=1}^{n} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{n}\sum_{A=1}^{N} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} = \sum_{i=1}^{n} \hat{h}(\mathbf{r}_i) \tag{1.14}$$

and a two-electron part

$$\hat{O}_2 = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{1.15}$$

The one-electron operator that appears on the RHS of Eq. (1.14),

$$\hat{h}(\mathbf{r}_i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{N} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|}, \tag{1.16}$$

is defined as the core Hamiltonian.

Using the Slater-Condon rules, [1, 2] the HF ground state energy can be expressed in terms of a set of one-electron and two-electron integrals:

$$E_{\text{HF}} = \sum_{i=1}^{n} \langle \chi_i | \hat{h} | \chi_i \rangle + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} [\langle \chi_i \chi_j | \chi_i \chi_j \rangle - \langle \chi_i \chi_j | \chi_j \chi_i \rangle] \tag{1.17a}$$

$$= \sum_{i=1}^{n} \langle i | \hat{h} | i \rangle + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} [\langle ij | ij \rangle - \langle ij | ji \rangle] \tag{1.17b}$$

where Eq. (1.17b) is a short-hand version of Eq. (1.17a). The integrals are represented using the Dirac bra-ket notation (also called the "physicist's notation" in electronic structure theory):

$$\langle \chi_i | \hat{h} | \chi_i \rangle = \int \chi_i^*(\mathbf{x}_1) \hat{h}(\mathbf{r}_1) \chi_i(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 \tag{1.18a}$$

$$\langle \chi_i \chi_j | \chi_i \chi_j \rangle = \int \int \chi_i^*(\mathbf{x}_1) \chi_j^*(\mathbf{x}_2) \frac{1}{r_{12}} \chi_i(\mathbf{x}_1) \chi_j(\mathbf{x}_2) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \tag{1.18b}$$

$$\langle \chi_i \chi_j | \chi_j \chi_i \rangle = \int \int \chi_i^*(\mathbf{x}_1) \chi_j^*(\mathbf{x}_2) \frac{1}{r_{12}} \chi_j(\mathbf{x}_1) \chi_i(\mathbf{x}_2) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2, \tag{1.18c}$$

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$. The integral given by Eq. (1.18b) is referred to as a Coulomb integral, which can be viewed as the Coulomb repulsion between the charge distributions associated with two spin orbitals ($|\chi_i(\mathbf{x}_1)|^2$ and $|\chi_j(\mathbf{x}_2)|^2$). The one given by Eq. (1.18c), on the other hand, is a pure quantum effect arising from the antisymmetrization of single-particle wave-function products, and it is called an exchange integral.

The Hartree-Fock energy functional given by Eq. (1.17a) can be minimized with respect to the spin orbitals ($\{\chi_i\}$) subject to the constraint that the spin orbitals are orthonormalized, i.e., $\langle \chi_i | \chi_j \rangle = \delta_{ij}$. This is shown to be equivalent to solving the following eigenvalue problem: [3]

$$\hat{f}(\mathbf{x}_1) \chi_i(\mathbf{x}_1) = \epsilon_i \chi_i(\mathbf{x}_1) \qquad i = 1, 2, \ldots, n \tag{1.19}$$

which is known as the Hartree-Fock equation. The Fock operator, $\hat{f}(\mathbf{x}_1)$, is another one-electron operator that plays an essential role in this thesis. It is expressed by

$$\hat{f}(\mathbf{x}_1) = \hat{h}(\mathbf{r}_1) + \sum_{j=1}^{n} \int \mathrm{d}\mathbf{x}_2 \chi_j^*(\mathbf{x}_2) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} (1 - \hat{P}_{12}) \chi_j(\mathbf{x}_2), \tag{1.20}$$

where $\hat{P}_{12}$ is a permutation operator that swaps $\mathbf{x}_1$ and $\mathbf{x}_2$: $\hat{P}_{12}\chi_j(\mathbf{x}_2)\chi_i(\mathbf{x}_1) = \chi_j(\mathbf{x}_1)\chi_i(\mathbf{x}_2)$.

According to the form of Eq. (1.20), a single electron (denoted by $\mathbf{x}_1$) "feels" the existence of other electrons through a mean-field potential. Therefore, under the Hartree-Fock approximation, the motions of different electrons are not correlated with each other. The electron correlation energy can thus be defined as the difference between the HF energy and the true ground state energy obtained by exactly solving the electronic Schrödinger equation:

$$E_{\text{corr}} = E_{\text{e},0} - E_{\text{HF}}, \tag{1.21}$$

which is a semi-negative definite quantity due to the variational feature of the HF theory. Electron correlation is extremely important for problems of interests in chemistry, including the accurate prediction of thermochemistry quantities (e.g. bond-breaking energies) and intermolecular interactions, and it is one of the central topics of modern electronic structure theory. One route to recover electron correlation energy is through wavefunction-based methods, such as configuration interaction (CI), coupled-cluster (CC) theory, many-body perturbation theory (MBPT), etc. These methods usually use the ground state wavefunction $|\Phi_0\rangle$ determined by the Hartree-Fock calculation as the reference, and incorporate excited determinants constructed by substituting occupied orbitals with virtual orbitals (they can both be obtained from an HF calculation, see Sec. 1.1.3). For example, the trial wavefunction given by a CI expansion can be expressed as

$$|\Phi_{\text{CI}}\rangle = c_0 |\Phi_0\rangle + \sum_{ia} c_i^a |\Phi_i^a\rangle + \sum_{ia,jb} c_{ij}^{ab} |\Phi_{ij}^{ab}\rangle + \dots \tag{1.22}$$

Note that many other wavefunction-based methods employ ansatz that correlates electrons explicitly, which were recently reviewed in Ref. 4.

Another approach to capture electron correlation is use of the density-functional approximation (DFA), which will be further discussed in Sec. 1.2.

1.1.3 Basics of an SCF Calculation

According to Eq. (1.20), the Fock operator depends on the spin orbitals, and therefore the Hartree-Fock equation needs to be solved through a self-consistent field (SCF) procedure, with a given basis set. This procedure is vital to the following chapters in this thesis since it also applies to a calculation using Kohn-Sham density functional theory [5], and we are going through its basics here.

To proceed, we first write down the Hartree-Fock equation for spatial orbitals $\{\psi_{i,\sigma}(\mathbf{r})\}$:

$$\hat{f}_\sigma(\mathbf{r}_1)\psi_{i,\sigma}(\mathbf{r_1}) = \epsilon_i \psi_{i,\sigma}(\mathbf{r_1}), \tag{1.23}$$

where

$$\hat{f}_\sigma(\mathbf{r}_1) = \hat{h}(\mathbf{r}_1) + \sum_{\sigma}^{\alpha,\beta}\sum_{j=1}^{n_\sigma} \int \mathrm{d}\mathbf{r}_2 \frac{|\psi_{j,\sigma}(\mathbf{r}_2)|^2}{|\mathbf{r}_1 - \mathbf{r}_2|} - \sum_{j=1}^{n_\sigma} \int \mathrm{d}\mathbf{r}_2 \psi_{j,\sigma}^*(\mathbf{r}_2) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \hat{P}_{12}\psi_{j,\sigma}(\mathbf{r}_2). \tag{1.24}$$

The Fock operator and the spatial orbital carry a spin index σ because the exchange operator acts differently for α and β electrons: based on Eq. (1.20), it is easy to show that exchange only occurs between same-spin electrons. For a closed-shell system, there is no spin polarization so two spin orbitals (of α and β spin) can share the same spatial component. This leads to a so-called restricted Hartree-Fock (RHF) calculation in which the same Fock operator applies to the α and β electrons. It is also possible to allow α and β electrons to use different sets of spatial orbitals, which is correspondingly named as unrestricted Hartree-Fock (UHF) and is particularly useful for open-shell systems (e.g. radicals).Without losing generality, we consider the spatial orbitals for the α electrons in the following discussion.

For molecular system calculations, each spatial orbital can be expanded in a finite atomic orbital (AO) basis set ($\{\omega_\nu(\mathbf{r})\}$):

$$\psi_{i,\alpha}(\mathbf{r}) = \sum_{\nu=1}^{N} \omega_\nu(\mathbf{r}) C_{\nu i,\alpha} \tag{1.25}$$

where N is the dimension of the employed basis set, and $\mathbf{C}_\sigma$ is referred to as the molecular orbital (MO) coefficient matrix. The AO basis functions are usually chosen to be atom-centered real functions and are *not* orthogonal to each other, so the overlap between two AO basis functions should be represented as $S_{\mu\nu} = \int \omega_\mu(\mathbf{r})\omega_\nu(\mathbf{r})\mathrm{d}\mathbf{r}$. By substituting Eq. (1.25) into (1.23), left-multiplying $\omega_\mu(\mathbf{r})$ and integrating over the 3-space, the Hartree-Fock equation is transformed into a generalized eigenvalue problem called the Roothaan equation: [6]

$$\sum_\nu F_{\mu\nu,\alpha} C_{\nu i,\alpha} = \sum_\nu S_{\mu\nu} C_{\nu i,\alpha} \epsilon_{i,\alpha}, \tag{1.26}$$

or

$$\mathbf{F}_\alpha \mathbf{C}_\alpha = \mathbf{S}\mathbf{C}_\alpha \boldsymbol{\epsilon}_\alpha, \tag{1.27}$$

where the matrix element of the AO Fock matrix can be expressed as

$$F_{\mu\nu,\alpha} = h_{\mu\nu} + \sum_{\sigma}^{\alpha,\beta} \sum_{j=1}^{n_\sigma} \langle \mu j_\sigma | \nu j_\sigma \rangle - \sum_{j=1}^{n_\alpha} \langle \mu j_\alpha | j_\alpha \nu \rangle . \tag{1.28}$$

The Fock matrix elements given by Eq. (1.28) can be expressed in a more compact form by introducing the one-particle density matrix (1PDM). The 3-space density corresponding to the Hartree-Fock solution ($|\Phi_0\rangle$) is given by

$$\begin{aligned} \rho(\mathbf{r}_1) &= N \int \mathrm{d}\sigma_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_3 \ldots \mathrm{d}\mathbf{x}_n |\Phi_0(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)|^2 \\ &= \sum_\sigma^{\alpha,\beta} \sum_{i=1}^{n_\sigma} |\psi_{i,\sigma}(\mathbf{r}_1)|^2. \end{aligned} \tag{1.29}$$

Let $\rho_\sigma(\mathbf{r}) = \sum_{i=1}^{n_\sigma} |\psi_{i,\sigma}(\mathbf{r})|^2$. By expanding the spatial orbitals with the AO basis, we have

$$\rho_\sigma(\mathbf{r}) = \sum_{i=1}^{n_\sigma} \sum_{\mu\nu} C_{\mu i,\sigma} C_{\nu i,\sigma} \omega_\mu(\mathbf{r}) \omega_\nu(\mathbf{r}) \tag{1.30}$$

$$= \sum_{\mu\nu} P_{\mu\nu,\sigma} \omega_\mu(\mathbf{r}) \omega_\nu(\mathbf{r}), \tag{1.31}$$

where the 1PDM (also loosely referred to as the density matrix) for spin σ (α or β) in the AO basis is defined as

$$P_{\mu\nu,\sigma} = \sum_{i=1}^{n_\sigma} C_{\mu i,\sigma} C_{\nu i,\sigma}. \tag{1.32}$$

Using the 1PDM, the Fock matrix for α electrons can be expressed as

$$F_{\mu\nu,\alpha} = h_{\mu\nu} + \sum_{\lambda\sigma} \langle \mu\lambda|\nu\sigma\rangle \left(P_{\lambda\sigma,\alpha} + P_{\lambda\sigma,\beta}\right) - \sum_{\lambda\sigma} \langle \mu\lambda|\sigma\nu\rangle P_{\lambda\sigma,\alpha}, \tag{1.33}$$

where the latter two terms define the Coulomb ($\mathbf{J}$) and exchange ($\mathbf{K}$) matrices:

$$J_{\mu\nu} = \sum_{\lambda\sigma} \langle \mu\lambda|\nu\sigma\rangle \left(P_{\lambda\sigma,\alpha} + P_{\lambda\sigma,\beta}\right), \quad K_{\mu\nu,\alpha} = \sum_{\lambda\sigma} \langle \mu\lambda|\sigma\nu\rangle P_{\lambda\sigma,\alpha}, \tag{1.34}$$

which are constructed by contracting the electron-repulsion integrals (ERIs) in the AO basis with the 1PDMs.

The Hartree-Fock energy given by Eq. (1.17b) can also be rewritten using the 1PDM and the Fock matrix:

$$E_{\mathrm{HF}} = \frac{1}{2}\left[\mathrm{Tr}[\mathbf{P}_\alpha(\mathbf{h} + \mathbf{F}_\alpha)] + \mathrm{Tr}[\mathbf{P}_\beta(\mathbf{h} + \mathbf{F}_\beta)]\right], \tag{1.35}$$

from which it is easy to derive that the Fock matrix is the derivative of the energy with respect to the density matrix: $\mathbf{F}_\sigma = \partial E_{\mathrm{HF}} / \partial \mathbf{P}_\sigma$.

Eqs. (1.33) and (1.35) give the Fock matrix and energy expressions routinely employed in SCF calculations in practice. In order to solve the Roothaan equation, the nonorthogonality of the AO basis also needs to be correctly handled, i.e., it is necessary to transform the generalized eigenvalue problem to an ordinary one. This is achieved by finding an AO basis orthogonalizer $\mathbf{X}$, such that $\mathbf{X}^T\mathbf{S}\mathbf{X} = \mathbf{I}$. Given $\mathbf{C} = \mathbf{X}\mathbf{C}'$ and $\mathbf{X}^T\mathbf{F}\mathbf{X} = \mathbf{F}'$, Eq. (1.27) can be rewritten as

$$\mathbf{F}'_\alpha \mathbf{C}'_\alpha = \mathbf{C}'_\alpha \boldsymbol{\epsilon}, \tag{1.36}$$

which can be solved by diagonalizing $\mathbf{F}'_\alpha$. The simplest way to construct $\mathbf{X}$ is through the symmetric orthogonalization scheme introduced by Löwdin,[7] in which $\mathbf{X} = \mathbf{S}^{-1/2}$.

Given a molecular system, an AO basis set, and an initial guess to this system's 1PDM ($\mathbf{P}_0$), its HF solution can be calculated with the following SCF procedure:

1. Calculate the matrices that are density-independent: $\mathbf{h}$, $\mathbf{S}$, $\mathbf{X}$.

2. Construct $\mathbf{J}$ and $\mathbf{K}$ with the guess 1PDM, and compute the initial Fock matrix based on Eq. (1.33).

3. Calculate $\mathbf{F}' = \mathbf{X}^T\mathbf{F}\mathbf{X}$; diagonalize $\mathbf{F}'$ to get $\mathbf{C}'$ and ϵ; obtain the updated MO coefficient matrix by $\mathbf{C} = \mathbf{X}\mathbf{C}'$.

4. Construct the updated density matrix ($\mathbf{P}$) based on Eq. (1.32).

5. Build $\mathbf{F}$ with the new density matrix.

6. Check for convergence based on a given criterion (e.g. sufficiently small change in $\mathbf{P}$):

   - If not converged, go back to Step 3 and iterate.
   - If converged, calculate the final HF energy based on Eq. (1.35) and terminate.

The procedure described above is sometimes referred to as the Roothaan-Hall algorithm. In order to accelerate convergence, modifications have been made to Step 3 to utilize an extrapolated Fock matrix, i.e., a linear combination of the most recent and previous Fock matrices, to generate the updated MOs. These more sophisticated techniques [8–11] are employed for the SCF calculations in this thesis.

The MO coefficient matrix ($\mathbf{C}_\sigma$) obtained from solving Eq. (1.27) has N_{orb} columns (N_{orb} is equal to the rank of the AO basis), while only the first n_σ columns corresponding to the *occupied orbitals* contribute to the HF energy. The remaining $N_{\text{orb}} - n_\sigma$ complementary vectors, as the "byproduct" of an SCF solution, are usually defined as the *virtual orbitals*. Therefore, besides being recast into a generalized eigenvalue problem, a Hartree-Fock calculation can also be regarded as an optimization of the partition between spaces spanned by occupied and virtual orbitals in the full AO space (the given "resolution of the identity"). Based on this view, an SCF calculation can be conducted by performing unitary rotations that mix the occupied and virtual orbitals. [12, 13] According to the scheme introduced in Ref. 13, these unitary rotations can be generated by using the exponential of an anti-Hermitian matrix:

$$\mathbf{C}^{(n)} = \mathbf{C}^{(n-1)} \exp \begin{pmatrix} \mathbf{0} & -\mathbf{\Delta}_{\text{vo}}^\dagger \\ \mathbf{\Delta}_{\text{vo}} & \mathbf{0} \end{pmatrix}, \tag{1.37}$$

and the SCF procedure is equivalent to a direct minimization problem to achieve the stationary condition $\partial E_{\text{HF}}/\partial \mathbf{\Delta} = \mathbf{0}$. The gradient of the HF energy with respect to the orbital rotations is closely related to the Fock matrix and the MOs that it is constructed from:

$$\frac{\partial E_{\text{HF}}}{\partial \mathbf{\Delta}_{\text{vo}}} = 2\mathbf{F}_{\text{ov}} = 2\mathbf{C}_{\text{o}}^T\mathbf{F}\mathbf{C}_{\text{v}}. \tag{1.38}$$

This alternative formulation of the SCF problem is also widely adopted in this thesis, and it inspires other analogous variational optimizations (see Chap. 2).

The cost of an SCF calculation depends on the size of the system (M) and the size of the employed AO basis set (represented by number of basis functions per atom, n). Considering

the Roothaan-Hall algorithm, there are two computationally significant steps in each SCF iteration: (i) constructing the Fock matrix (especially **J** and **K**) from the given density; (ii) diagonalizing the Fock matrix to update **C** (and **P**). With the conventional algorithm, the former scales as $\mathcal{O}(M^2)$ (the asymptotic behavior arising from the sparsity of AO function pairs) and $\mathcal{O}(n^4)$, and the latter scales as $\mathcal{O}(M^3)$ and $\mathcal{O}(n^3)$. Despite the lower asymptotic scaling with respect to the system size, building the Fock matrix usually dominates the computational cost for routinely studied molecular systems because of the high demands associated with the evaluation of ERIs. In the past two decades, many techniques have been introduced to reduce the cost associated with SCF calculations, by lowering the scaling with respect to M or n, and a brief review of these methods can be found in Sec. 2.1. The objective of the method introduced in Chap. 2 is also to make SCF calculations employing large basis sets more feasible in terms of the associated computational costs.

1.2 Density Functional Theory

1.2.1 Basic Principles

Density functional theory (DFT) is the main workhorse for today's electronic structure calculations because it provides a way to incorporate many electron correlation effects with a similar cost to that of a Hartree-Fock calculation. It uses the 3-space electron density, $\rho(\mathbf{r})$, as the basic variable. The foundation of DFT is the two theorems proven by Hohenberg and Kohn (HK) in 1964. [14] The first Hohenberg-Kohn theorem (HKI) asserts that *there is a one-to-one mapping between the ground state electron density* $\rho(\mathbf{r})$ *and the external potential* $v_{\text{ext}}(\mathbf{r})$. By solving the electronic Schrödinger equation with the given $v_{\text{ext}}(\mathbf{r})$, the resulting ground state wavefunction is also a functional of $\rho(\mathbf{r})$:

$$\rho(\mathbf{r}) \rightarrow v_{\text{ext}}(\mathbf{r}) \rightarrow \Psi_0[\rho(\mathbf{r})]. \tag{1.39}$$

For a molecular system in vacuum, $\hat{V}_{\text{ext}}$ is simply $\hat{V}_{\text{ne}}$. Based on HKI, the ground state energy of the system can be expressed as a functional of $\rho(\mathbf{r})$:

$$\begin{aligned} E[\rho(\mathbf{r})] &= \left\langle \Psi_0[\rho(\mathbf{r})] \middle| \hat{T} + \hat{V}_{\text{ee}} + \hat{V}_{\text{ne}} \middle| \Psi_0[\rho(\mathbf{r})] \right\rangle \\ &= T[\rho(\mathbf{r})] + V_{\text{ee}}[\rho(\mathbf{r})] + \int v_{\text{ne}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} \qquad (1.40a) \\ &= T[\rho(\mathbf{r})] + J[\rho(\mathbf{r})] + K[\rho(\mathbf{r})] + \int v_{\text{ne}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r}, \qquad (1.40b) \end{aligned}$$

where in Eq. (1.40b) the V_{ee} term is further split into the classical Coulomb repulsion

$$J[\rho(\mathbf{r})] = \frac{1}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{1.41}$$

and a non-classical contribution ($K[\rho]$) that incorporates effects such as exchange and correlation. The second HK theorem (HKII) ensures that *the exact ground state density minimizes the total energy of the system*, which validates the variational search for the optimal $\rho(\mathbf{r})$ and the corresponding $E[\rho]$.

The two HK theorems provide an alternative approach (sometimes referred to as HK-DFT or orbital-free (OF)-DFT [15]) to the exact ground state energy of the system, which has a very simple form as $\rho(\mathbf{r})$ is a function of only three independent variables (x, y, z). However, it faces two major obstacles in practical application. The first obstacle arises from the requirement of v-representability, i.e., the trial density $\tilde{\rho}(\mathbf{r})$ must come from integrating a valid ground state wavefunction associated with a valid $v_{\text{ext}}(\mathbf{r})$, while the HK theorems provide no recipe for constructing $\tilde{\rho}(\mathbf{r})$ of a given system. The second obstacle is the limitation of accuracy yielded by OF-DFT. Among the four terms on the RHS of Eq. (1.40b), the exact form for the kinetic energy functional ($T[\rho]$) and the non-classical electron-electron interaction ($K[\rho]$) are unknown, where the former is usually of the same magnitude as the total energy. Therefore, to find an accurate approximation for the kinetic energy functional is crucial to the applicability of OF-DFT to chemical systems, which, however, is a very difficult task. Despite many recent efforts [16–19], the successful applications of OF-DFT are limited to systems whose electronic structure is not too far away from the uniform electron gas (UEG), such as main-group metals/alloys [20–22] and some semi-conductors [23].

The Kohn-Sham formulation of density functional theory (KS-DFT) [5] circumvents the obstacle of inaccurate treatment of kinetic energy by re-introducing the concept of orbitals. It describes a fictitious reference system consisting of n non-interacting electrons that reside in n spin orbitals (referred to as KS orbitals), which, by using Eq. (1.29), yield the same 3-space density as the true ground state wavefunction. As the exact solution to the Schrödinger equation for this non-interacting system is simply the Slater determinant constituted with these orbitals, the kinetic energy of this system, T_s, can be evaluated in the same way as in the HF theory:

$$T_s = \sum_{i=1}^{n} \left\langle \chi_i \left| -\frac{1}{2}\nabla_i^2 \right| \chi_i \right\rangle = \sum_{\sigma}^{\alpha,\beta} \sum_{i=1}^{n_\sigma} \left\langle \psi_{i,\sigma} \left| -\frac{1}{2}\nabla_i^2 \right| \psi_{i,\sigma} \right\rangle, \tag{1.42}$$

where the subscript "s" indicates that this kinetic energy is calculated from a Slater determinant. As this Slater determinant is obtained by solving a special Schrödinger equation whose $\hat{V}_{\text{ne}}$ maps to $\rho(\mathbf{r})$ and $\hat{V}_{\text{ee}} = 0$, T_s is also a functional of $\rho(\mathbf{r})$. Then, by introducing a new term called the *exchange-correlation* (XC) functional, Eq. (1.40b) can be rewritten as

$$E[\rho] = T_s[\rho] + \int v_{\text{ne}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} + J[\rho] + E_{\text{xc}}[\rho], \tag{1.43}$$

where

$$E_{\text{xc}}[\rho] = (T[\rho] - T_s[\rho]) + (V_{\text{ee}}[\rho] - J[\rho]) \tag{1.44}$$

is a combination of the deviation of T_s from the exact kinetic energy and the non-classical part of the electron-electron interaction.

The KS formalism of DFT is an exact theory, although $E_{\text{xc}}[\rho]$, the only term in Eq. (1.43) whose exact form is unknown, has to be evaluated approximately in practice. The last three decades have seen almost countless efforts on improving the approximation to the XC functional (which will be briefly summarized in Sec. 1.2.2), making KS-DFT today's most popular electronic structure method. The key to the success of KS-DFT is that the kinetic energy of the reference system is usually not too far away from the exact kinetic energy, and as a result, the inexactly treated term E_{xc} is usually one magnitude smaller than the total energy. Therefore, it is more feasible to come up with an accurate approximation to E_{xc} than in the OF-DFT case where the target is the entire $T[\rho]$. From now on, "DFT" and "KS-DFT" are used synonymously in this thesis unless otherwise specified.

The KS energy functional can be variationally optimized. By differentiating Eq. (1.43) with respect to $\psi_{i,\sigma}(\mathbf{r})$ and utilizing Eq. (1.29) , it can be recast into a set of operator equations similar to Eq. (1.23):

$$\hat{f}_{\sigma}^{\text{KS}}(\mathbf{r})\psi_{i,\sigma}(\mathbf{r}) = \epsilon_i \psi_{i,\sigma}(\mathbf{r}), \tag{1.45}$$

where

$$\hat{f}_{\sigma}^{\text{KS}}(\mathbf{r}) = -\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|} + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r} + v_{\text{xc},\sigma}(\mathbf{r}) \tag{1.46}$$

defines the effective one-electron Hamiltonian (Fock operator) for KS-DFT. The last term on the RHS of Eq. (1.46) (the XC potential) is the functional derivative of E_{xc} with respect to the electron density: $v_{\text{xc},\sigma}(\mathbf{r}) = \delta E_{\text{xc}}/\delta\rho_{\sigma}(\mathbf{r})$.

Eq. (1.45) can be transformed into a generalized eigenvalue problem by expanding the KS orbitals with an AO basis:

$$\mathbf{F}_{\sigma}^{\text{KS}}\mathbf{C}_{\sigma} = \mathbf{S}\mathbf{C}_{\sigma}\boldsymbol{\epsilon}_{\sigma}, \tag{1.47}$$

where $\mathbf{F}_{\sigma}^{\text{KS}} = \mathbf{h} + \mathbf{J} + \mathbf{V}_{\text{xc},\sigma}$ is the KS Fock matrix. It can be solved in the same way as the Roothaan equation via the SCF procedure introduced in Sec. 1.1.3. The total electronic energy of the system can also be represented using the 1PDMs ($\mathbf{P}_{\alpha}, \mathbf{P}_{\beta}$):

$$E_{\text{KS}} = \text{Tr}[(\mathbf{P}_{\alpha}+\mathbf{P}_{\beta})\mathbf{h}] + \text{Tr}[(\mathbf{P}_{\alpha}+\mathbf{P}_{\beta})\mathbf{J}] + E_{\text{xc}}[\mathbf{P}_{\alpha}, \mathbf{P}_{\beta}]. \tag{1.48}$$

Note that one difference from the Hartree-Fock case is that E_{xc} and $\mathbf{V}_{\text{xc}}$ usually cannot be calculated analytically. Instead, they are evaluated through numerical integrations performed on a 3-dimensional real space grid.

1.2.2 Overview of Modern Density Functionals

Even though the exact form of $E_{\text{xc}}[\rho]$ is elusive, the past thirty years have seen an explosion of interest in KS-DFT, particularly in ways to approximate the XC functional and to improve the performance of DFT for non-covalent interactions. These advances are worth mentioning here, as they motivate and validate the works in this thesis, and we refer the reader to Refs. [24–29] for reviews of the recent developments.

Although the density functional approximation (DFA) is known to be not systematically improvable (which is the main reason for the plethora of XC functionals ever proposed), a well-accepted way to introduce the hierarchy of DFT is through Perdew's "Jacob's ladder" [30] as shown in Fig. 1.1. The ground below this ladder is the so-called "Hartree's world", where electrons "feel" each other only through the classical Coulomb repulsion and there is no exchange or correlation effect. One step up the ladder indicates that a new ingredient is added to the DFA, and a well-designed functional at a higher rung, in principle, should outperform the lower-rung functionals in terms of the general accuracy. The ultimate goal (the "heaven") is to attain chemical accuracy for various types of systems with DFA.

The exchange (E_{x}) and correlation (E_{c}) components are usually separately treated in a DFA. While exchange only occurs between same-spin electrons, correlation is not subject to the same restriction and thus E_{c} has both same-spin and an opposite-spin components. For convenience, we first write down the most general form for the exchange and correlation functionals considered in this thesis:

$$E_{\mathrm{x}} = \sum_{\sigma}^{\alpha,\beta} \int e_{\mathrm{x},\sigma} \left[\rho_\sigma, |\nabla\rho_\sigma|, \tau_\sigma\right] \mathrm{d}\mathbf{r} \tag{1.49a}$$

$$E_{\mathrm{c}} = \int \rho(\mathbf{r}) \epsilon_{\mathrm{c}} \left[\rho_\alpha, \rho_\beta, |\nabla\rho_\alpha|, |\nabla\rho_\beta|, \nabla\rho_\alpha \cdot \nabla\rho_\beta, \tau_\alpha, \tau_\beta\right] \mathrm{d}\mathbf{r}, \tag{1.49b}$$

where $e_{\mathrm{x},\sigma}$ is defined as the exchange energy for α or β electrons per volume, ϵ_{c} as the correlation energy per electron, and the other variables are introduced below.

The rung 1 functionals on the Jacob's ladder employ the local-spin density approximation (LSDA), in which E_{xc} relies on ρ_α and ρ_β only. The analytical form of the LSDA exchange functional can be derived ($e_{\mathrm{x},\sigma}(\mathbf{r}) = -C_x\rho(\mathbf{r})^{4/3}$) [31], which is known as the Slater functional. The LSDA correlation functionals such as VWN5 [32] and PW92 [33], on the other hand, were obtained by fitting to quantum Monte-Carlo data. As the simplest DFA, LSDA provides an exact description for the uniform electron gas (UEG) and is widely used in the condensed-phase physics community. However, it has very poor performance for describing chemical systems (e.g., LSDA often overestimates bond energies by > 100 kJ/mol) primarily because electron density is fast varying and thus far from being uniform.

The second-rung functionals introduce the 3-space gradient of electron density ($\nabla\rho_\sigma$, or its dimensionless form $x_\sigma = |\nabla\rho_\sigma|/\rho_\sigma^{4/3}$) as a new variable to account for the above-mentioned inhomogeneity. Functionals of this type are referred to as generalized-gradient approximation (GGA) functionals. The use of density gradient information offers a significant improvement over LSDA (for properties such as atomization energies of molecules, the error is reduced by roughly an order of magnitude), and it is arguably the first useful class of functionals for chemistry. Popular GGA exchange-correlation functionals include BLYP (B88 [34] + LYP [35]), PBE [36], revPBE [37], and Grimme's dispersion-corrected B97-D [38] functional.

Besides density gradients, an additional variable (τ_σ) is employed by the third-rung functionals. τ_σ is the kinetic energy density of the KS reference system and can be evaluated from the KS orbitals: $\tau_\sigma = \sum_i^{n_\sigma} |\nabla\psi_{i,\sigma}|^2$. This variable, albeit local, is found to be capable

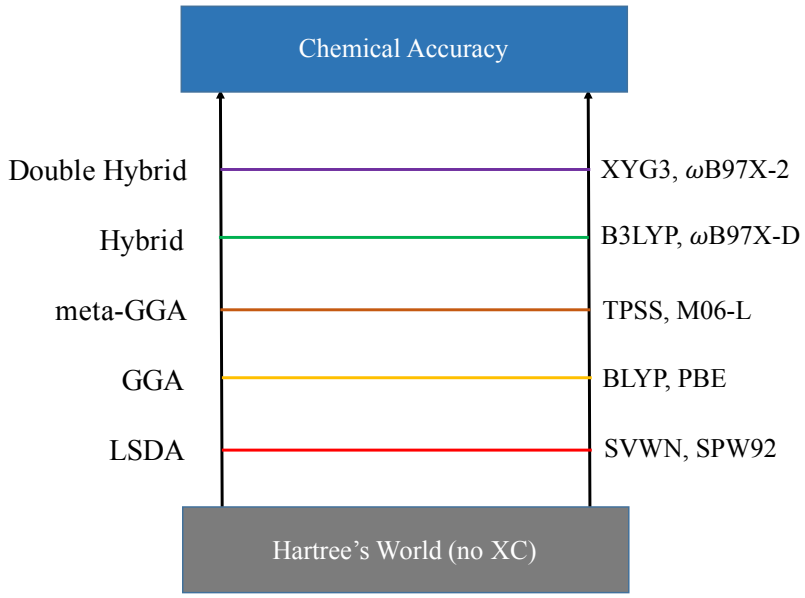

Figure 1.1: Perdew's "Jacob's ladder" for density functionals. The name of the functional category corresponding to each rung is denoted on the left, and two examples are given on the right for each category (rung).

of detecting non-local exchange holes in systems [39] and thus is expected to offer a potential improvement over the GGAs. This class of functionals is referred to as the meta-GGAs, and the most well-known examples include TPSS (non-empirical) [40] and M06-L (empirical) [41]. The meta-GGA functionals, in principle, should outperform the GGAs as an extra variable is utilized, and the improvement is most noticeable on thermochemistry properties such as atomization energies, barrier heights, etc., according to the results in Ref. 25.

The functionals on the first three rungs are sometimes referred to as local or semi-local functionals because the employed variables are functions of a single 3-space coordinate ($\mathbf{r}$). These functionals suffer from the notorious issue of self-interaction error (SIE), i.e., $\hat{J}(\mathbf{r})$ and $\hat{v}_{\text{xc}}(\mathbf{r})$ fail to exactly cancel each other for a one-electron KS orbital. The SIE renders semi-local functionals inaccurate for properties such as barrier heights, isomerization energies, HOMO-LUMO gap, etc. The simplest way to alleviate this problem is to replace a fraction of the semi-local exchange functional with exact (Hartree-Fock) exchange globally, yielding the global hybrid (GH) GGA or GH meta-GGA functionals that belong to the fourth rung of the ladder. The GH functionals have become very popular since the early 1990s mainly because of the improved accuracy that they provide in general. One prototypical and arguably the most popular GH functional is Becke's 3-parameter functional with the LYP correlation, known as B3LYP [42]:

$$E_{\text{xc}}^{\text{B3LYP}} = (1 - a_{\text{x}} - c_{\text{x}})E_{\text{x}}^{\text{Slater}} + c_{\text{x}}E_{\text{x}}^{\text{HF}} + a_{\text{x}}E_{\text{x}}^{\text{B88}} + (1 - a_{\text{c}})E_{\text{c}}^{\text{VWN}} + a_{\text{c}}E_{\text{c}}^{\text{LYP}}, \tag{1.50}$$

for which the fraction of HF exchange (c_{x}) is 0.20 (globally). Other popular global hybrids

include the non-empirical PBE0 (GH GGA, 25% HFX) [43] and the highly parameterized M06-2X (GH meta-GGA, 54% HFX) [44] functionals.

The fourth-rung of the Jacob's ladder also covers the range-separated hybrid (RSH) functionals. The basic idea underlying RSH is to increase the fraction of HF exchange for well-separated electrons (up to 100%), whereby the correct long-range behavior of the exchange energy can be reproduced. The range-separation is realized by partitioning the electron-electron Coulomb operator into a short-range (SR) and a long-range (LR) part [45, 46], and their range is controlled by the ω parameter:

$$\frac{1}{r_{12}} = \frac{1-\mathrm{erf}(\omega r_{12})}{r_{12}} + \frac{\mathrm{erf}(\omega r_{12})}{r_{12}}, \tag{1.51}$$

where $\mathrm{erf}(x)$ is the error function which goes to 1 when $x \to \infty$. Given the SR and LR Coulomb operators, the exchange energy of an RSH functional can be most generally expressed as:

$$E_{\mathrm{x}}^{\mathrm{RSH}} = c_{\mathrm{x,sr}} E_{\mathrm{x,sr}}^{\mathrm{HF}} + (1-c_{\mathrm{x,sr}}) E_{\mathrm{x,sr}}^{\mathrm{DFT}} + c_{\mathrm{x,lr}} E_{\mathrm{x,lr}}^{\mathrm{HF}} + (1-c_{\mathrm{x,lr}}) E_{\mathrm{x,lr}}^{\mathrm{DFT}}, \tag{1.52}$$

where $E_{\mathrm{x,sr}}^{\mathrm{HF}}$ and $E_{\mathrm{x,lr}}^{\mathrm{HF}}$ are calculated with the SR and LR version of the $1/r_{12}$ operators, respectively, and the SR and LR DFT exchanges are also correspondingly "dressed". RSH can be regarded as an improvement over GH as the range-separation is physically motivated, and the RSH functionals, in principle, are able to better address the problems caused by SIE (such as exaggerated electron delocalization). Well-known examples of RSH functionals include CAM-B3LYP [47], ωB97/ωB97X [48] , and ωB97X-D [49].

The rung 5 functionals make use of correlated wavefunction theory (e.g. MP2 [50], RPA [51]) to recover the missing dynamical correlation effect, and therefore virtual KS orbitals are needed in the energy evaluation. They are referred to as the double-hybrid functionals [52], and a few representative examples include B2PLYP [53], ωB97X-2 [54], and XYG3 [55]. Since these functionals are not employed for any calculations in this thesis, we will not discuss them further.

Climbing the Jacob's ladder from the ground to rung 4, however, does not overcome the problem of missing long-range dispersion force in mean-field calculations. This shortcoming inhibits the use of KS-DFT for weakly bound systems. A remedy for this is to introduce dispersion corrections in DFT calculations, and the past decade has seen a significant number of methods proposed for this purpose[27–29]. The most straightforward approach is to add a C_n-potential ($n = 6, 8, \dots$), which is an analytical function of internuclear distances, to the DFT energy. The most popular examples are Grimme's DFT-D methods [38, 56, 57], in which the C_n coefficients are pre-tabulated. The now widely used DFT-D3 scheme[57] employs a long look-up table containing $C_{n,ij}$ for each atom pair (ij) among the 1–94 elements, and the dispersion correction is given by

$$E_{\mathrm{D3}} = -\sum_{i<j} \sum_{n=6,8} s_n \frac{C_{n,ij}}{R_{ij}^n} f_{\mathrm{damp},n}(R_{ij}), \tag{1.53}$$

where s_n is a functional-dependent global scaling factor, and $f_{\mathrm{damp},n}$ is a damping function (which has many flavors available [57–61]) that switches off this correction when the interatomic distance (R_{ij}) is small. Besides DFT-D, other approaches that compute the C_n coefficients based on the actual electronic structure such as the exchange-dipole moment (XDM) [62, 63] and Tkatchenko-Scheffler (TS) [64] models also belong to this category.

Another approach to incorporate dispersion in DFT calculations is using the non-local correlation (NLC) functionals. These functionals share the following general form:

$$E_{\mathrm{NLC}}[\rho] = \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \rho(\mathbf{r}) \Phi(\mathbf{r}, \mathbf{r}') \rho(\mathbf{r}'), \tag{1.54}$$

where $\Phi(\mathbf{r}, \mathbf{r}')$ is the NLC kernel. The evaluation of E_{NLC} requires a double numerical integration so this approach is more computationally demanding than the C_n potentials. Nonetheless, this approach has received more attentions recently because of its *ab initio* feature, the successful practice of coupling NLC with XC functionals, and the advances in efficient computations [65, 66]. Two most notable classes of NLC are the *ab initio* vdW-DFs by Langreth and co-workers [67–69] and the semi-empirical Vydrov-Van Voorhis kernels (VV09 [70] and VV10 [71]).

Most recently, three combinatorially designed empirical functionals, namely ωB97X-V (RSH GGA) [72], B97M-V (semi-local meta-GGA) [73], and ωB97M-V (RSH meta-GGA) [74], were developed by Mardirossian and Head-Gordon at Berkeley. These functionals employ the power series of inhomogeneity variables (essentially $|\nabla\rho_\sigma|$ and τ_σ) and use a "survival-of-the-fittest" strategy to choose the linear parameters to fit. In this sense, these functionals are minimally parameterized and thus are expected to be maximally transferable. The above-mentioned VV10 [71] NLC functional is employed to account for long-range dispersion. Based on a comprehensive benchmark on a gigantic molecular database (MGCDB84) that comprises roughly 5000 data points [25], these functionals are the most accurate ones in general among the functionals belonging to the same category. Their accuracy for non-covalent interactions is particularly remarkable, as demonstrated by this and other recent benchmarks [75–81]. These functionals' capability of accuracy, to a large extent, motivates the work of this thesis, and they are directly employed for most of the calculations reported here.

1.3 Intermolecular Interaction

1.3.1 Definition and Significance

Intermolecular interactions refer to the repulsive or attractive forces that act between molecules or other microscopic objects that are composed of nuclei and electrons (e.g. atoms, ions, etc.). The strengths of these interactions considered in this thesis cover a broad range, from a few tenths of kJ/mol for rare gas dimers to hundreds of kJ/mol for strong donor-acceptor complexes concerning equilibrium binding energies. While a majority of these

interactions are weaker than intramolecular forces (chemical bonds), there is no clear-cut distinction between them in terms of magnitude.

Intermolecular interactions are ubiquitous in nature and play an important role in determining the macroscopic properties and microscopic structures of substances. The discovery of these interactions was related to the deviation of real gas behaviors from those predicted by the ideal gas law. For instance, a non-zero second virial coefficient, as a correction to the ideal gas law for the pressure-density relation, is a direct consequence of two-body interactions between gas molecules. The existence of intermolecular interactions is a necessary condition for the formation of molecular liquids and solids, and their thermodynamic properties such as boiling/melting points and densities are altered by these interactions. These interactions also govern the folding of proteins, the pairing of bases during the replication of DNA, and the binding of substrates with the active sites of enzymes, and are thus crucial to biological processes.

Understanding and tuning intermolecular interactions is of great importance in many branches of today's chemical research. In the design of gas-storage materials, such as metal-organic frameworks (MOFs) [82, 83], it is useful to unravel the nature of interactions between gas molecules and either metal centers or linkers. Similarly, the mechanism behind protein-ligand binding is crucial for the design of new drugs that can precisely act on targets (e.g. a specific site on a protein). For chemical reactions, there are countless examples where the interactions between reactant molecules, between substrates and catalysts, and between reactive species and environment exert significant influence on the reaction mechanisms. In these scenarios, knowledge of intermolecular interactions can guide chemical modifications that lead to desired consequences (high gas-storage capacity, selective drug-acceptor binding, low reaction barrier, etc.). The manipulation of intermolecular binding also lays the foundation for the concept of supramolecular chemistry [84, 85], which has many applications in self-assembled materials, catalysis, molecular recognition, molecular devices, etc.

Intermolecular potentials can be extracted from experimental approaches such as molecular-beam scattering, although such measurements are limited to interactions between small molecules in the gas phase. Nevertheless, there is an abundance of effects due to intermolecular interactions that can be experimentally characterized. For example, the so-called solvatochromic effect (the color of a compound changes upon solvation) [86] reveals how the ground and excited states of a chromophore interact differently with the solvent environment. The structural and vibrational properties of a binding moiety can also be altered upon the formation of an intermolecular complex, several examples of which are studied in Chap. 3.

Theoretical approaches, on the other hand, are powerful tools for understanding and modeling intermolecular interactions. Recent development of electronic structure theory has made it possible to compute accurate intermolecular potentials without inputs from experimental data, and analysis tools are available to quantify the contributions from different physical components to binding. Besides quantum mechanics (QM)-based approaches, empirical models such as molecular mechanics (MM) force fields employ parameterized analytical functions to incorporate intermolecular interactions in molecular simulations, which require much lower computational costs than *ab initio* methods and are able to provide useful

insight into chemical processes in condensed-phase systems. An overview on these aspects is provided in the rest of Sec. 1.3 as well as in Sec. 1.4.

1.3.2 Ab initio Calculation of Interaction Energies

The so-called "supermolecular" approach is employed in this work to compute *interaction energies* (ΔE_{int}) between molecules, i.e., for a system comprising N_F fragments, the interaction energy is defined as the difference between the energy of the supersystem (E_{full}) and the sum of isolated fragment energies (E_F):

$$\Delta E_{\text{int}} = E_{\text{full}} - \sum_{F=1}^{N_F} E_F. \tag{1.55}$$

The total interaction energy can thus be determined by performing $N_F + 1$ single-point calculations, with a given theoretical *model chemistry*, i.e., the combination of a theory (WFT or DFT) and a basis set. Noted that the fragment energies in Eq. (1.55) are evaluated at the geometry of the complex. If we further take the geometry distortion (GD) effect into account, i.e., the energy penalty for fully relaxed monomers to change into their geometries in the complex: $\Delta E_{\text{gd}} = \sum_F (E_F - E_F^{(0)})$, we obtain the intermolecular *binding energy*:

$$\Delta E_{\text{bind}} = \Delta E_{\text{int}} + \Delta E_{\text{gd}}. \tag{1.56}$$

While ΔE_{bind} and ΔE_{int} are often used synonymously in literature, they refer to different quantities in this thesis.

One technical issue associated with the supermolecular approach is the basis set superposition error (BSSE). It originates from the use of atom-centered basis sets, i.e., the calculations for each monomer and the supersystem are not performed in the same Hilbert space. Due to BSSE, Eq. (1.55) or (1.56) often yields systematically overbound results when incomplete basis sets are employed. A comprehensive investigation of this effect associated with Gaussian basis sets for DFT-calculated dimer interaction energies was recently performed by Witte et al.[87]. For non-covalently interacting fragments, BSSE can be removed by using the Boys-Bernardi counterpoise (CP) correction [88], in which the energy of each fragment is also evaluated in the supersystem basis (giving energies denoted as $E_F(\text{sys})$):

$$\Delta E_{\text{int}}^{\text{CP}} = E_{\text{full}} - \sum_{F=1}^{N_F} E_F(\text{sys}), \tag{1.57}$$

and the BSSE is correspondingly defined as

$$\text{BSSE} = \sum_{F=1}^{N_F} E_F - E_F(\text{sys}). \tag{1.58}$$

Most recently, geometry-based corrections for BSSE were proposed in the DFT community in order to overcome the cumbersomeness of the Boys-Bernardi approach for molecular clusters, and to address the challenge of intramolecular BSSE. These approaches employ empirical functions of nuclear positions and are parameterized for specific basis sets. Examples can be found in Refs. [89–91].

It should be noted that besides the supermolecular (subtractive) approach, intermolecular interaction energies can also be calculated with additive formalisms. Symmetry-adapted perturbation theory (SAPT) [92] is a prominent example of this class of methods. For a system consisting of two fragments A and B, SAPT partitions the electronic Hamiltonian as

$$\hat{H} = \hat{H}_0^{(A)} + \hat{H}_0^{B} + \xi\hat{W}^{(A)} + \eta\hat{W}^{(B)} + \zeta\hat{V}^{(AB)}, \tag{1.59}$$

where $\hat{H}_0^{(A)}$ refers to the mean-field Hamiltonian for fragment A, $\hat{W}^{(A)} = \hat{H}^{(A)} - \hat{H}_0^{(A)}$ is the corresponding Møller-Plesset [50] fluctuation potential, and $\hat{V}^{(AB)}$ is the Coulomb interaction between two fragments:

$$\hat{V}^{(AB)} = \sum_{i,I\in A}\sum_{j,J\in B}\left(\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \frac{Z_I}{|\mathbf{r}_j - \mathbf{R}_I|} - \frac{Z_J}{|\mathbf{r}_i - \mathbf{R}_J|} + \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}\right). \tag{1.60}$$

Using symmetry-adapted Rayleigh-Schrödinger (SRS) perturbation theory, the total interaction can be written as an expansion in orders of $\hat{V}^{(AB)}$, $\hat{W}^{(A)}$, and $\hat{W}^{(B)}$ (represented by n, k, and l, respectively):

$$E_{\text{int}}^{\text{SAPT}} = \sum_{n=1}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}(E_{\text{pol}}^{(nkl)} + E_{\text{exch}}^{(nkl)}). \tag{1.61}$$

The "pol" terms come from the so-called polarization expansion (which is evaluated with the Hartree product of monomer wavefunctions), while the associated "exch" terms originate from the enforcement of antisymmetry of supersystem wavefunction with regard to the exchange between electrons on different fragments. This series provides a hierarchy of *ab initio* methods for calculating intermolecular interactions. Meanwhile, as at least the low-order terms in the SAPT expansion are associated with intuitive physical meanings, it provides a decomposition of the total interaction energy. Although the present thesis is not concerned with the development and application of SAPT, it will be further discussed in Sec. (1.3.3) and the following chapters as an important alternative approach to energy decomposition analysis (EDA).

1.3.3 Physical Components and Energy Decomposition Analysis

While the total strength of intermolecular interaction can be calculated via the approaches introduced above, sometimes this quantity solely is insufficient to provide a useful guide for practical applications. To achieve a desired objective in chemical synthesis, experimental means such as introducing a substituent group are commonly utilized to modify the

strength of binding between two moieties. Knowledge of the physical components of the given interaction is often necessary before one can rationalize the choice of a substituent group. Descriptive concepts based on both the underlying physics and chemical intuition, such as permanent electrostatics, polarization, London dispersion, donor-acceptor interaction, etc., are useful in these scenarios, since the general behavior of each of these components are relatively well-established. For example, if the interaction is dominated by charge-transfer effects, it is possible for one to enhance this interaction by introducing an electron-withdrawing group on the acceptor to make it more electron-deficient, or an electron-donating group on the donor to increase its Lewis basicity. The same modification, on the other hand, may not be as effective if this interaction is dominated by other effects such as permanent electrostatics or dispersion. Quantifying the physical components of intermolecular interactions is also meaningful for the development of empirical additive models such as MM force fields. The non-bonded (intermolecular) terms in a force field such as Coulomb interactions and repulsive/attractive van der Waals (vdW) potentials are designed based on their distinct asymptotic behaviors. Therefore, a parameterization based on the magnitude of individual components at different intermolecular separations, in principle, is able to produce a more accurate and transferable force field than, say, a less physically justified form. [93]

Physical components of intermolecular interactions considered in this thesis are summarized as follows:

- **Permanent Electrostatics**
  For well-separated fragments, permanent electrostatics can be defined as the classical Coulomb interaction between the intrinsic charge distribution of each fragment. As a multipole expansion provides an accurate approximation for long-range Coulomb interactions, its asymptotic distance dependence can be determined by the leading multipole moment of each fragment. For example, the electrostatic interaction for the water-Na^+ complex decays as R^{-2} (charge-dipole) in the long range. In the short range, a multipole expansion is incapable of accurately describing permanent electrostatics as fragment charge distributions overlap with each other. The charge penetration effect [94] emerges and renders the actual permanent electrostatic interaction more favorable than that given by multipole expansion. This effect, however, diminishes exponentially with the increase of the intermolecular distance.

- **Pauli Repulsion and Exchange**
  Pauli repulsion is a quantum mechanical effect, and its name originates from the Pauli exclusion principle that *no two electrons can occupy the same spatial and spin coordinate*. This repulsive force arises from the antisymmetrization of fragment wavefunctions, and the energy increase mostly resides in the kinetic energy of the supersystem. The rise in kinetic energy due to the spatial confinement imposed by other molecules nearby can be rationalized based on the uncertainty principle. Another effect arising from the antisymmetry of electronic wavefunction is exchange, which, contrasting with Pauli repulsion, is an attractive effect. The easiest way to understand this term is through the derivation of

the Hartree-Fock theory (see Sec. 1.1.2). Because of the common origin of the above two components, they are often considered together as a composite term in many theoretical models, and they both decay exponentially with the increase of intermolecular separation.

- **Polarization**
  In classical electromagnetism, the term polarization is used to describe the response of a dielectric medium to an external electric field. In the context of an intermolecular interaction, polarization refers to the *intramolecular relaxation* in the presence of other molecules, which is often accompanied by fragment charge redistributions and energy lowering of the supersystem. The rearrangement of fragment charges can be further converted into concepts such as induced dipoles via a multipole expansion. In the classical picture, the magnitude of an induced dipole is proportional to the strength of the external field and the polarizability of the medium. This picture is useful for describing molecular polarization when fragments are well-separated. Still using the water-Na^+ example, its polarization energy has an R^{-4} long-range distance dependence as it is dominated by the charge-induced dipole interaction. Two important features of polarization in molecular systems are (i) the mutual polarization renders it not pairwise additive, as the induced multipole moments further interact with each other; (ii) in the short range, polarization is modulated by the requirement of wavefunction antisymmetry so it differs from pure electrical polarization.

- **Dispersion**
  The dispersion interaction is another purely non-classical effect which was first introduced by London in the 1930s [95]. It arises from the correlated motion of electrons belonging to different fragments, and thus it cannot be properly described by mean-field approaches such as Hartree-Fock and conventional density functionals. On the other hand, it is incorporated in interaction energies calculated by dispersion-corrected DFT [27, 28] or correlated wavefunction methods (MP2 [50], CCSD(T) [96], ...). Dispersion is always a favorable effect. For weakly bound systems whose permanent multipole interactions are unfavorable (e.g. π-stacking systems) or zero (e.g. rare gas molecules), dispersion is the dominant binding force. In the asymptotic regime, dispersion decays as R^{-6}, while it is not well-defined when molecules are in close contact as there is no clear cut between dispersion and other short-range correlation effects.

- **Charge Transfer**
  In contrast to polarization, charge transfer (CT) refers to the stabilization effect attributed to the *intermolecular relaxations* of the supersystem wavefunction. Under the MO picture, it can be viewed as mixing the empty orbitals of other fragments into the occupied orbitals of a given fragment. These intermolecular relaxations mainly lower the kinetic energy of the system as electrons on each fragment become more delocalized, and are often accompanied with electron population changes on each fragment. Concepts such as electron *donor* and *acceptor* (based on the decrease/increase in electron population) are commonly used to describe CT between a pair of molecules. Like Pauli repulsion, CT is

also a short-range effect and its magnitude decays exponentially with the intermolecular distance.

It should be noted that these physical components are not quantum mechanical observables and do not have unique definitions (especially in the overlapping regime). However, they are extremely useful concepts for describing intermolecular interactions based on the reasons given above. Energy decomposition analysis (EDA) is a category of methods that bridges these concepts and electronic structure calculations, i.e., it partitions the total interaction energy into contributions from the aforementioned components such as permanent electrostatics, Pauli repulsion, dispersion, etc. While many EDA schemes are available (see the Introduction of Ref. [97] for a summary of these methods and Refs. [98, 99] for comprehensive reviews), two types of approaches are of particular importance in our opinion: (i) SAPT-based methods; (ii) variational EDA schemes combined with modern density functionals.

As noted in Sec. 1.3.2, SAPT provides an avenue to the total interaction energy through a perturbative expansion (Eq. (1.61)), in which each term is associated with a physical meaning. For instance, truncating the expansion at the second-order for the intermolecular potential $\hat{V}$ ($n = 2$) and zeroth-order for $\hat{W}_A, \hat{W}_B$ yields the SAPT0 energy:

$$E_{\text{int}}^{\text{SAPT0}} = E_{\text{elec}}^{(1)} + E_{\text{exch}}^{(1)} + E_{\text{ind}}^{(2)} + E_{\text{exch-ind}}^{(2)} + E_{\text{disp}}^{(2)} + E_{\text{exch-disp}}^{(2)}, \tag{1.62}$$

which is decomposed into contributions from electrostatics ($E_{\text{elec}}^{(1)}$), exchange ($E_{\text{exch}}^{(1)}$), induction ($E_{\text{ind}}^{(2)} + E_{\text{exch-ind}}^{(2)}$) and dispersion ($E_{\text{disp}}^{(2)} + E_{\text{exch-disp}}^{(2)}$). Note that under the SAPT terminology, "exchange" accounts for the Pauli repulsion effect introduced above, and "induction" encompasses both polarization and charge-transfer effects. More accurate interaction energies can be obtained by truncating Eq. (1.61) at a higher order (e.g. SAPT2+) [100], which, however, is computationally demanding for large systems.

The SAPT-based energy decomposition schemes face several challenges: (i) their applications are mostly limited to dimer systems, as the perturbative expansion has a cumbersome form for systems comprising more than two fragments; (ii) the accuracy of the total interaction energy relies on the convergence of perturbative expansion, which turns out to be difficult for systems involving strong induction effects (e.g., when ionic species is involved) [75]; (iii) there is no natural way to separate polarization and charge transfer within the framework of SAPT although several approaches have been suggested [101–103], and this will be further discussed in Chap. 4.

Variational EDA methods, on the other hand, aim to decompose the interaction energy calculated via the supermolecular approach (Eq. (1.55)). Intermediate states that are variational upper bounds to the full SCF solution are involved in these methods, and the energy components are defined as the energy differences between these intermediate states, or between the intermediate and the full SCF states. These methods originate from the Kitaura-Morokuma (KM)-EDA [104, 105] and the Ziegler-Rauk EDA (also known as the extended transition state (ETS) method) [106, 107] that were proposed in the 1970s, and

more recent approaches such as the Bickelhaupt-Barends EDA [108, 109], block-localized wavefunction (BLW)-EDA [110–112], density-based EDA (DEDA) [113, 114], and absolutely localized molecular orbital (ALMO)-EDA [97, 115, 116] all belong to this class. In the early days, variational EDA was performed with inaccurate model chemistries (e.g. HF/4-31G in Ref. 105) so they were regarded as tools for qualitative analysis, while the significantly improved accuracy of modern density functionals for non-covalent interactions validates use of these methods for quantitative purposes.

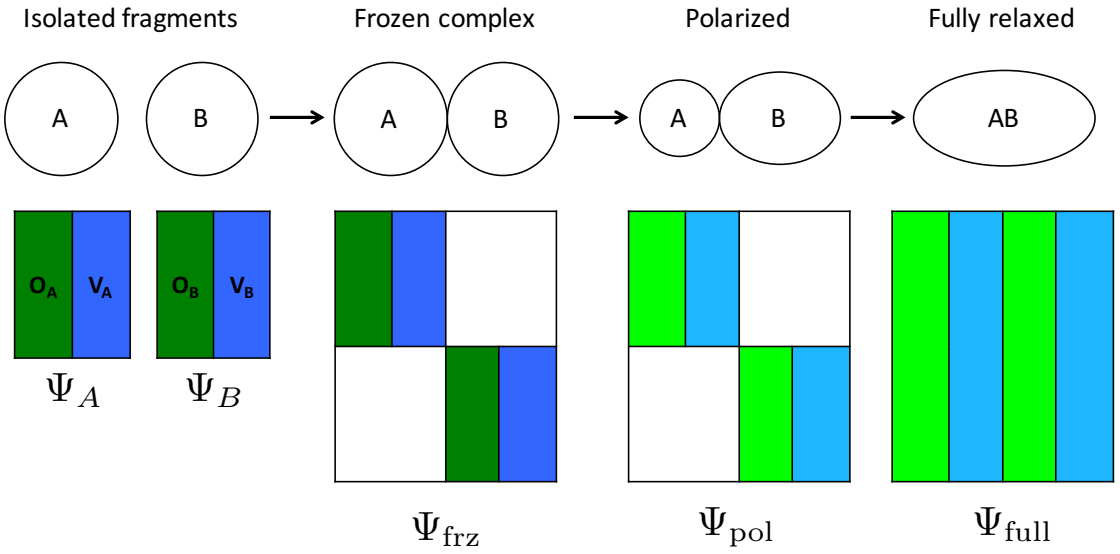

Figure 1.2: Illustration of the ALMO-EDA procedure with the intermediate states denoted.

A major part of this thesis is concerned with the development and application of the ALMO-EDA method, whose original version was proposed by Head-Gordon and co-workers in 2007 [115] and further extended to open-shell systems in 2013 [116]. It decomposes the total interaction energy into three components: frozen interaction, polarization, and charge transfer, and its main feature is the use of a variationally optimized fragment-block-diagonal MO coefficient matrix for the polarized yet CT-forbidden intermediate state. The procedure is illustrated in Fig. 1.2 and will be discussed in detail in Chaps. 3–6. Recent advances in this method include addressing the basis set limit issue for the separation of polarization and charge transfer [117] and enabling the further decomposition of the frozen term into contributions from permanent electrostatics, Pauli repulsion, and dispersion [118], which define the second-generation ALMO-EDA method [97]. Meanwhile, an "adiabatic" version of the ALMO-EDA was proposed in order to characterize the effects of intermolecular interactions on molecular properties [119], which is covered in Chap. 3 of this thesis. ALMO-EDA has also been extended to post-HF methods such as MP2 [120, 121] and scenarios where excited state molecules are involved [122].

1.4 Potential Energy Surface for Molecular Simulation

Molecular simulation is a powerful tool for probing properties of bulk systems and investigating processes occurring in either gas phase or condensed phase. Under the BO approximation (Sec. 1.1.1), nuclei move on a potential energy surface (PES) that corresponds to a given electronic state. The configuration space of nuclear coordinates can be sampled by performing Monte Carlo (MC) or molecular dynamics (MD) simulations. The former only requires energy evaluations on a PES, whereas the latter usually also requires force calculations to propagate the system forwards in time. Amongst a number of aspects relevant for the validity of a molecular simulation (ensembles, sampling techniques, etc.) [123], the quality of the employed PES is of crucial importance, as it determines the accuracy of the results extracted from the simulation. Here we briefly introduce several PESs that are commonly employed in atomistic molecular simulations, as they are the subjects of the studies in Chaps. 5 and 6.

According to Sec. 1.1.1, a fully *ab initio* PES can be constructed by solving the electronic Schrödinger equation at each nuclear configuration. Propagating the system with Newtonian equations (fully classical) on an *ab initio* PES gives rise to the simulation technique usually known as BOMD.[124] BOMD requires an analytical gradient (force) calculation at each time step, which, in principle, can be done with any well-defined electronic structure method. For example, the nuclear gradient of KS-DFT can be represented as

$$E^x_{\mathrm{KS}} = \mathbf{P}\cdot\mathbf{h}^x + \frac{1}{2}\,\mathbf{P}\cdot\mathbf{II}^x\cdot\mathbf{P} - (\mathbf{PFP})\cdot\mathbf{S}^x + E^x_{\mathrm{xc}}, \tag{1.63}$$

where the superscript "x" indicates a differentiation with respect to a component of nuclear position, "**II**" represents the AO ERIs, and the other matrices have the same meaning as in Sec. 1.1. The four terms in Eq. (1.63) represent the contributions from the core Hamiltonian, AO ERIs, AO overlap matrix, and the XC energy. Provided that an atom-centered Gaussian basis set is employed, the differentiation takes place through both the operators and basis functions, giving rise to the so-called Hellman-Feynman force [125] and Pulay force [126], respectively. For example, the gradient of **h** consists of three terms:

$$h^x_{\mu\nu} = \langle\omega_\mu|\hat{h}^x|\omega_\nu\rangle + \langle\omega^x_\mu|\hat{h}|\omega_\nu\rangle + \langle\omega_\mu|\hat{h}|\omega^x_\nu\rangle\,, \tag{1.64}$$

where the last two terms correspond to the Pulay force. The evaluation of the analytical gradient of E_{xc} was elaborated in Ref. 127, and we are not presenting more details here. Note that the analytical gradient of the KS-DFT energy lays the foundation for the "adiabatic" EDA method introduced in Chap. 3.

While BOMD and its approximate variant, Car-Parrinello molecular dynamics (CPMD)[128], have led to many important findings thanks to the accuracy of the *ab initio* PES, their routine use for large-scale condensed-phase systems is still unaffordable. This is because an enormous number of electronic structure calculations are required to obtain

converged results, due to the vast number of degrees of freedom to sample. Therefore, in practice one often needs to resort to PESs whose computational demand is more tractable. Molecular mechanical (MM) force fields based fixed-charge models are at the lower end of the "cost-accuracy spectrum" of available PESs. In these models, simple analytical functions of nuclear coordinates are utilized to capture intra- and intermolecular forces, and its basic form has been available since the 1960s [129]:

$$\begin{aligned} U = &\sum_{b}^{\text{bonds}} k_b(b-b_0)^2 + \sum_{\theta}^{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\phi}^{\text{torsions}} k_\phi[1+\cos(n\phi+\delta)] \\ &+ \sum_{i<j} q_i q_j / r_{ij} + \sum_{i<j} 4\epsilon_{ij}\left[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}\right]. \end{aligned} \tag{1.65}$$

The first three terms describe the energy changes due to the small fluctuations of bond lengths, bond angles and dihedrals from their values at the optimal geometry. They are referred to as the "bonded terms" or "valence terms" and connectivity information is required to evaluate them. The last two "non-bonded" terms account for the long-range electrostatics and vdW interactions, which are described by Coulomb interaction between point charges and the well-known Lennard-Jones (12-6) potential, respectively. The point charges are usually located on atom centers, although occasionally off-center sites are used [130]. The potential energy (and its analytical gradient) is calculated using pre-tabulated parameters based on atom types, and examples of well-known parameterizations include the Amber [131], CHARMM [132], OPLS [133], and GROMOS [134] force fields.

While these fixed-charge models have been widely applied to the simulations of condensed-phase systems such as biomolecules [135, 136], they have several evident discrepancies from the correct physics concerning intermolecular interactions. First, their intermolecular potentials (the last two terms in Eq. (1.65)) are pairwise additive. As a result, they are not able to describe the many-body effects in condensed-phase systems such as the cooperativity or anti-cooperativity of hydrogen bonds. The second shortcoming is their poor description of the short-range effects such as charge penetration, Pauli repulsion, and charge transfer, as they are not explicitly included in Eq. (1.65). These effects, however, are of great importance to the microscopic structures and properties of condensed-phase systems even at ambient conditions [137]. To address the first weakness, explicit polarization has been introduced to MM force fields via models based on fluctuating charges, Drude oscillators, or inducible dipoles (see Ref. 138 for a review of these approaches). The AMOEBA force field [139–141] is a prominent example of a polarizable model. It is investigated and employed in the last two chapters of this thesis. The general form of the full AMOEBA potential comprises the following terms:

$$\begin{aligned} U = &U_{\text{bond}} + U_{\text{angle}} + U_{\text{bnd-ang}} + U_{\text{oop}} + U_{\text{torsion}} \\ &+ U_{\text{perm}} + U_{\text{pol}} + U_{\text{vdw}}, \end{aligned} \tag{1.66}$$

where the last three terms correspond to the non-bonded potential. The most important feature of AMOEBA is that an isotropic polarizability is assigned to each atomic site and

the induced dipoles are determined self-consistently, whereby it captures the many-body polarization effect. Other features include use of higher atomic multipoles (up to quadrupoles) for permanent electrostatics and a buffered 14-7 vdW potential. It is worth mentioning that revisions of AMOEBA's intermolecular potential form and parameters is an on-going effort, including the addition of explicit terms for charge penetration [142, 143] and charge transfer [144], as well as the revision of polarization [145] and vdW models [146]. The details about the non-bonded terms in AMOEBA and the recent efforts on including additional short-range effects in MM force fields are further elaborated in Chap. 5.

Conventional MM force fields are incapable of describing chemical reactions in the condensed phase that involve bond breaking or changes in chemical valence, such as enzyme reactions, alkane cracking in zeolites, etc. The hybrid QM/MM methods [147–152] emerge as a natural compromise between accuracy and efficiency for the simulation of such reactive chemical processes. The idea is to treat a small key part of the entire system (e.g. the enzyme active site and substrate) with electronic structure methods, and the less important surroundings (e.g. the side chains and solvent molecules) with less accurate and cheaper MM. The most popular formalism of QM/MM at present is arguably the so-called "electrostatic embedding" (EE) approach, in which the continuous QM density interacts with the Coulomb potential of the MM charges:

$$E_{\text{elec}}^{\text{QM-MM}} = \int \mathrm{d}\mathbf{r}\ \rho(\mathbf{r}) V_{\text{chg}}(\mathbf{r}), \quad V_{\text{chg}}(\mathbf{r}) = \sum_{i=1}^{N_{\text{chg}}} \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|}. \tag{1.67}$$

$\hat{V}_{\text{chg}}$ enters the electronic structure calculation as an additional term in the core-Hamiltonian. In this way, the QM subsystem is polarized by MM charges through the SCF procedure. The steric effect (Pauli repulsion) and London dispersion between the QM and MM subsystems, however, are usually computed at the MM level:

$$E_{\text{vdw}}^{\text{QM-MM}} = \sum_{i=1}^{n_{\text{QM}}} \sum_{j=1}^{n_{\text{MM}}} 4\epsilon_{ij} \left[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6} \right]. \tag{1.68}$$

EE with fixed-charge models neglects the response of the MM subsystem to the electric field imposed by the QM wavefunction and nuclei. This effect, however, is of particular importance in scenarios where the electronic structure of the QM part varies drastically during the simulated process (e.g. an electron-transfer reaction). Therefore, it is desirable to further incorporate MM polarization in QM/MM calculations, yielding mutually polarizable QM/MM (sometimes called "polarizable embedding" (PE)). The QM/AMOEBA model described in Chap. 6 belongs to this category. Many other approaches to PE, such as methods based on effective fragment potentials (EFP) [153–155], are also briefly reviewed in that chapter. The main challenge faced by today's QM/MM methods is the inadequate accuracy for interactions crossing the QM/MM boundary, where the errors arise from the physically incorrect treatment of short-range effects such as charge penetration and Pauli repulsion. As a consequence, very large QM regions (10^2–10^3 atoms) were found to be necessary to

achieve the desired accuracy [156–159], which, to some extent, defeats the purpose of using QM/MM.

Besides the three categories of PESs (*ab initio*, MM force field, and QM/MM) discussed above, there are many other options available on the "cost-accuracy spectrum". Two noteworthy examples are the semi-empirical density functional tight-binding (DFTB) method [160–162], and the "MB-pol" potential (many-body potential energy functions parameterized for specific species) by Paesani and co-workers[163–165]. As these approaches are neither investigated nor employed in the following chapters, we will not further introduce them.

1.5 Outline

Motivated by the significantly improved accuracy provided by KS-DFT and the demand for advanced potential energy surfaces for molecular simulations, this thesis reports several recent advances in schemes aiming to accelerate DFT calculations and in DFT-based energy decomposition analysis (EDA) methods. In Chapter 2, a new SCF scheme using a minimal adaptive basis constructed on-the-fly is introduced, which is a promising method to attain the desired accuracy of modern density functionals at the basis set limit with reduced computational costs. Chapter 3 introduces the "adiabatic EDA" concept and reformulates the ALMO-EDA method accordingly, and demonstrates how it is useful for probing the observable effects of intermolecular interactions. Chapter 4 is a comprehensive discussion of the definition of charge transfer in EDA methods, which, to a large extent, relies on the adiabatic EDA scheme developed in Chapter 3. Chapters 5 and 6 aim to apply the second-generation ALMO-EDA method to the development of PESs for multi-scale simulation. Chapter 5 assesses the AMOEBA force field on simple ion-water interactions against the ALMO-EDA results and provides a detailed analysis of the strengths and weaknesses of the current form of the AMOEBA force field. Chapter 6 proposes and implements a mutually polarizable QM/MM model using the AMOEBA force field and provides a diagnosis for the interactions across the QM/MM boundary using the same type of methodology as in Chapter 5. A slightly more detailed synopsis of each chapter follows.

Chapter 2

Recently developed density functionals have good accuracy for both thermochemistry (TC) and non-covalent interactions (NC) if very large atomic orbital basis sets are used. To approach the basis set limit with potentially lower computational cost, a new self-consistent field (SCF) scheme is presented that employs minimal adaptive basis (MAB) functions. The MAB functions are optimized on each atomic site by minimizing a surrogate function. High accuracy is obtained by applying a perturbative correction (PC) to the MAB calculation, similar to dual basis approaches. Compared to exact SCF results, using this MAB-SCF (PC) approach with the same large target basis set produces < 0.15 kcal/mol root-mean-square deviations (RMSDs) for most of the tested TC datasets, and < 0.1 kcal/mol for most of

the NC datasets. The performance of density functionals near the basis set limit can be even better reproduced. With further improvement to its implementation, MAB-SCF (PC) is a promising lower-cost substitute for conventional large-basis calculations as a method to approach the basis set limit of modern density functionals. This work [166] has been published in *The Journal of Chemical Physics*.

Chapter 3

Energy decomposition analysis (EDA) of electronic structure calculations has facilitated quantitative understanding of diverse intermolecular interactions. Nevertheless, such analyses are usually performed at a single geometry and thus decompose a "single-point" interaction energy. As a result, the influence of the physically meaningful EDA components on the molecular structure and other properties are not directly obtained. To address this gap, the absolutely localized molecular orbital (ALMO)-EDA is reformulated in an adiabatic picture, where the frozen, polarization, and charge transfer energy contributions are defined as energy differences between the stationary points on different potential energy surfaces (PESs), which are accessed by geometry optimizations at the frozen, polarized and fully relaxed levels of density functional theory (DFT). Other molecular properties such as vibrational frequencies can thus be obtained at the stationary points on each PES. We apply the adiabatic ALMO-EDA to different configurations of the water dimer, the water-Cl^- and water-Mg^{2+}/Ca^{2+} complexes, metallocenes (Fe^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}), and the ammonia-borane complex. This method appears to be very useful for unraveling how physical effects such as polarization and charge transfer modulate changes in molecular properties induced by intermolecular interactions. As an example of the insight obtained, we find that a linear hydrogen bond geometry for the water dimer is preferred even without the presence of polarization and charge transfer, while the red shift in the OH stretch frequency is primarily a charge transfer effect; by contrast, a near-linear geometry for the water-chloride hydrogen bond is achieved only when charge transfer is allowed. This work [119] has been published in *Physical Chemistry Chemical Physics*.

Chapter 4

Charge-transfer (CT) is an important binding force in the formation of intermolecular complexes, and there have been a variety of theoretical models proposed to quantify this effect. These approaches, which typically rely on a definition of a "CT-free" state based on a partition of the system, sometimes yield significantly different results for a given intermolecular complex. Two widely used definitions of the "CT-free" state, the absolutely localized molecular orbitals (ALMO) method (where only on-fragment orbital mixings are permitted) and the constrained density functional theory (CDFT) approach (where fragment electron populations are fixed), are carefully examined in this work. Natural bond orbital (NBO) and the regularized symmetry-adapted perturbation theory (SAPT) are also briefly considered. Results for the ALMO and CDFT definitions of CT are compared on a broad range of model

systems, including hydrogen-bonding systems, borane complexes, metal-carbonyl complexes, and complexes formed by water and metal cations. For most of these systems, CDFT yields a much smaller equilibrium CT energy compared to that given by the ALMO-based definition. This is mainly because the CDFT population constraint does not fully inhibit CT, which means that the CDFT "CT-free" state is in fact CT-contaminated. Examples of this contamination include (i) matching forward and backward donation (e.g. formic acid dimer), and (ii) unidirectional CT without changing fragment populations. The magnitude of the latter effect is quantified in systems such as the water dimer by employing a 3-space density constraint in addition to the orbital constraint. Furthermore, by means of the adiabatic EDA, it is shown that several observable effects of CT, such as the "pyramidalization" of the planar BH_3 molecule upon the complexation with Lewis bases, already appear on the "CT-free" CDFT surface. These results reveal the essential distinctions between the ALMO and CDFT definitions of CT, and suggest that the former is more consistent with accepted understanding of the role of CT in intermolecular binding.

Chapter 5

AMOEBA is a molecular mechanics force field that addresses some of the shortcomings of a fixed partial charge model, by including permanent atomic point multipoles through quadrupoles, as well as many-body polarization through the use of point inducible dipoles. In this work we investigate how well AMOEBA formulates its non-bonded interactions, and how it implicitly incorporates quantum mechanical effects such as charge penetration (CP) and charge transfer (CT), for water-water and water-ion interactions. We find that AMOEBA's total interaction energies as a function of distance and over angular scans for the water dimer and for a range of water-monovalent cations agree well with an advanced density functional theory (DFT) model, whereas the water-halides and water-divalent cations show significant disagreement with the DFT result, especially in the compressed region when the two fragments overlap. We use a second generation energy decomposition analysis (EDA) scheme based on absolutely localized molecular orbitals (ALMOs) to show that in the best cases AMOEBA relies on cancellation of errors by softening of the van der Waals (vdW) wall to balance permanent electrostatics that are too unfavorable, thereby compensating for the missing CP effect. CT, as another important stabilizing effect not explicitly accounted for in AMOEBA, is also found to be incorporated by the softened vdW interaction. For the water-halides and water-divalent cations, this compensatory approach is not as well executed by AMOEBA over all distances and angles, wherein permanent electrostatics remains too unfavorable and polarization is over-damped in the former while overestimated in the latter. We conclude that the DFT-based EDA approach can help refine a next generation AMOEBA model that either realizes a better cancellation of errors for problematic cases like those illustrated here, or to guide the parameterization of explicit functional forms for short-range contributions from CP and/or CT. This work [167] has been published in *Journal of Chemical Theory and Computation*.

Chapter 6

The importance of incorporating solvent polarization effects into the modeling of solvation processes has been well-recognized, and therefore a new generation of hybrid quantum mechanics/molecular mechanics (QM/MM) approaches that accounts for this effect is desirable. We present a fully self-consistent, mutually polarizable QM/MM scheme using the AMOEBA force field, in which the total energy of the system is variationally minimized with respect to both the QM electronic density and the MM induced dipoles. This QM/AMOEBA model is implemented through the Q-Chem/LibEFP code interface and then applied to the evaluation of solute-solvent interaction energies for various systems ranging from the water dimer to neutral and ionic solutes (NH_3, NH_4^+, CN^-) surrounded by increasing numbers of water molecules (up to 100). In order to analyze the resulting interaction energies, we also utilize an energy decomposition analysis (EDA) scheme which identifies contributions from permanent electrostatics, polarization and van der Waals (vdW) interaction for the interaction between the QM solute and the solvent molecules described by AMOEBA. This facilitates a component-wise comparison against full QM calculations where the corresponding energy components are obtained via a modified version of the absolutely localized molecular orbitals (ALMO)-EDA. The results show that the present QM/AMOEBA model can yield reasonable solute-solvent interaction energies for neutral and cationic species, while further scrutiny reveals that this accuracy highly relies on the delicate balance between insufficiently favorable permanent electrostatics and softened vdW interaction. For anionic solutes where the charge penetration effect becomes more pronounced, the QM/MM interface turns out to be unbalanced. These results are consistent with and further elucidate our findings in a previous study using a slightly different QM/AMOEBA model (Ref. 168). The implications of these results for further refinement of this model are also discussed. This work [169] has been published in *Journal of Chemical Theory and Computation*.

Chapter 2

Development of the MAB-SCF Scheme

2.1 Introduction

Kohn-Sham density functional theory [5, 14, 170] (KS-DFT) has become the most widespread electronic-structure method because of its reasonable balance between accuracy and computational cost. Functionals using the generalized gradient approximation (GGA) [36, 171] are usually regarded as the simplest that can give acceptable accuracy for chemistry. To overcome the plague of self-interaction error, new variables have been introduced, leading to meta-GGA [40, 172, 173], global hybrid (GH) [42], and range-separated hybrid (RSH) [45, 46]functionals. At the same time, a variety of models have been developed to account for van der Waals (vdW) interactions within DFT [27, 28], including the empirical DFT-D methods [38, 56, 57] and nonlocal correlation (NLC) functionals (e.g. vdw-DF-10 [69], VV10 [71]). Most recently, two combinatorially designed functionals were developed: ωB97X-V[72] (RSH+VV10) and B97M-V [73](meta-GGA+VV10), which demonstrated impressive accuracy on both thermochemistry (TC) and non-covalent interactions (NC), with an accessible complete basis set (CBS) limit, and low grid sensitivity.

With finite atomic orbital (AO) basis sets [127], one prerequisite for attaining such accuracy is to approach the CBS limit. This issue has been carefully investigated [174–177], but is often neglected in practical applications, as exemplified by the prevalence of the "B3LYP/6-31G(d)" model chemistry. A basis set of *at least* triple-ζ and *preferably* quadruple-ζ quality is often required by hybrid functionals (e.g. B3LYP[34, 35, 42]) to obtain adequately converged thermochemistry results. Even for the semi-local B97M-V functional, the acceptable alternatives to aug-cc-pVQZ [178, 179] (which almost represents the CBS limit) are still of least triple-ζ quality. Turning to the evaluation of NCs, a similar study on the A24 [180] and S66 [181, 182] complexes indicates that augmented triple-ζ basis sets (e.g. aug-cc-pVTZ, def2-TZVPD [183]) are in general required by B97M-V to properly converge the binding energies. Their double-ζ counterparts (e.g. aug-cc-pVDZ, def2-SVPD) should only be carefully

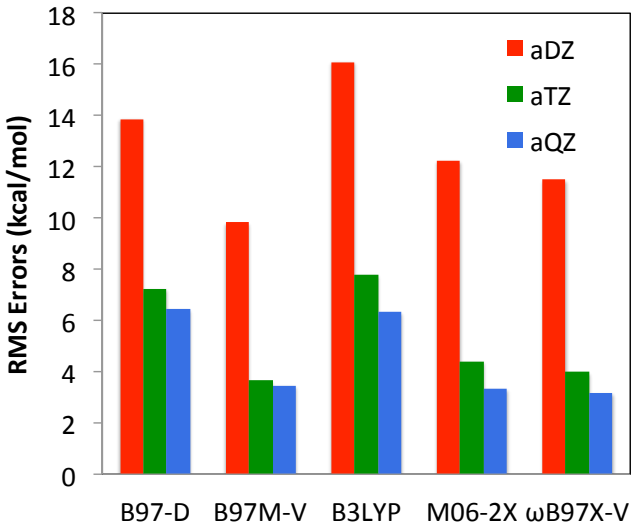

Figure 2.1: Basis set dependence of the root-mean-square (RMS) errors for G2 atomization energies against the reference values from Ref. 186. Data for five functionals are collected: B97-D (GGA), B97M-V (meta-GGA), B3LYP (GH), M06-2X (GH) [44], and ωB97X-V (RSH). Dunning's augmented correlation-consistent basis sets (D, T, Q) are used to systematically approach the CBS limit.

used with counterpoise (CP) corrections [184]. For ωB97X-V, the requirement on basis set qualities might be even higher due to the slower basis set convergence of functionals that contain exact exchange.

Figure 2.1 demonstrates the basis set convergence of several modern density functionals in terms of their RMSDs for the G2 set [185](atomization energies of 148 neutral molecules, whose reference values are taken from Ref. 186). Apart from B97M-V, using aQZ instead of aTZ for the other four functionals reduces their RMSDs by 0.8–1.4 kcal/mol, including the semi-local B97-D [38] functional (GGA). Using aDZ yields very poor accuracy (usually over 10 kcal/mol RMSDs) for all these functionals, which defeats the purpose of using state-of-the-art density functionals. One way to tackle this problem is by directly training a functional in a small basis, such as the EDF1 functional [187], which was parameterized at the 6-31+G(d) level. By relying on cancellation between functional error and basis set error, the transferability of these methods are often limited, and further empirical corrections seem necessary to achieve adequate accuracy for relative energies [90].

Each self-consistent field (SCF) cycle of a KS-DFT calculation involves two computationally demanding steps: (1) Fock matrix construction with a given density, and (2) Fock matrix diagonalization to update the density. For fixed system size, the computational cost of the Fock build scales as $\mathcal{O}(n^4)$ with respect to the basis size (n) when conventional AO algorithms are used, and the cost of the diagonalization step scales as $\mathcal{O}(n^3)$. This steep cost increase inhibits large basis sets (e.g. those of quadruple-ζ size) from being routinely employed in DFT calculations. It should be noted that the scaling of cost vs. basis size

is largely independent of the development of linear scaling (with system size) Fock matrix build algorithms [188–195] and many diagonalization replacements [196, 197]. Moreover, near-complete basis sets are not favored by linear-scaling algorithms, especially when diffuse functions are included, since matrix element sparsity is diminished and the overlap matrix starts to be ill-conditioned (note that optimization of Gaussian basis sets up to triple-ζ quality with much reduced condition numbers while maintaining condensed phase accuracy has been reported [198]), which in turn destroys the sparsity of the density matrix [199, 200].

One successful strategy to make large basis KS-DFT calculations more tractable is to compute the full Coulomb ($\mathbf{J}$) and exchange ($\mathbf{K}$) matrices more efficiently by approximating two-electron repulsion integrals (ERIs) with the aid of auxiliary basis functions or grid points. The resolution-of-the-identity (RI) method [201–203] expands the product of AO function pairs with a preoptimized auxiliary basis. RI algorithms do not improve the system-size scaling unless local fit regions are applied [204, 205], but they reduce the basis set size scaling from $\mathcal{O}(n^4)$ to $\mathcal{O}(n^3)$. Therefore, state-of-the-art RI algorithms (e.g. MARI-J [206], PARI-K [207], occ-RI-K [208]) can speed up Fock matrix constructions in large basis sets significantly for small- to medium-sized systems, while retaining numerical accuracy. The diagonalization step is unaffected.

A second successful approach to accelerating large basis calculations is to perform the iterative SCF procedures in a primary (small) basis and then approximate the secondary (target) basis results by utilizing perturbation theory. This idea was introduced for post-SCF methods (e.g. MP2) [209, 210], and was then developed for SCF methods, including the dual-basis SCF (DB-SCF) method developed by Head-Gordon and co-workers[211–215], and the "Hartree-Fock/Density Functional Perturbation Corrections" (HFPC/DFPC) scheme proposed by Deng and Gill[216–218]. With a careful choice of primary/secondary basis set pairing, these methods can provide satisfactory accuracy for both TC [211, 217, 218] and NC [215] with significantly reduced computational costs (roughly 10 times faster), although system-size scaling remains unchanged. One limitation is the need to develop and validate the basis set pairings [215], which determines accuracy and speedup. Furthermore, as the secondary basis approaches the CBS limit, the size of the primary basis needed to achieve a given accuracy also increases: for instance the optimized primary basis for cc-pVQZ is roughly of cc-pVTZ size [212].

A related approach is the use of small adaptive basis sets. The idea of encoding chemical environment information into atomic/quasiatomic basis functions to understand chemical bonding dates back to early tools [219–224], as well as some more contemporary methods [225–230]. Apart from interpretive purposes, the merits of utilizing small adaptive bases in KS-DFT calculations have been recognized with the development of fast (especially linear-scaling) SCF algorithms, leading to renewed interest in the concept of "polarized atomic orbitals" (PAOs), first put forward by Adams in the 1960s [231, 232]. These adaptive sets usually have very tiny (often minimal) spans, which leads to vastly fewer variational degrees of freedom. In addition, an adaptive basis constructed with spatial confinement contributes to a well-conditioned overlap matrix, which is a property favored by $\mathcal{O}(N)$ scaling methods.

The Adams PAO scheme treats atoms in a molecule as fragments and solves projected

equations self-consistently on each of them, which is similar, in spirit, to projected SCF methods using fragment-localized, non-orthogonal molecular orbitals (MOs) to evaluate intermolecular interactions (SCF-MI) [116, 233, 234]. In practice this scheme only works for weakly interacting atoms (e.g. rare-gas clusters) or ionic compounds (e.g. LiH, NaCl) [235]. Later, the PAO approach was recast to form a minimal atom-centered adaptive basis as an atom-blocked contraction of the secondary basis functions on each atom [236]. The molecular energy is minimized simultaneously with respect to the atom-blocked contraction coefficients and the density matrix in the adaptive basis [236]. The PAO-SCF energy can be improved using perturbation theory [237], similar to the dual basis approaches discussed above. The minimal rank of the PAO basis and its atomic locality makes it promising for linear-scaling algorithms [200], but the "double" optimization problem is challenging and often causes convergence problems.

Significant progress on tractable adaptive basis schemes for KS-DFT has been made in the condensed matter physics community. Similar to the aforementioned PAO approach, Ozaki and Kino [238, 239] and others [240, 241] used numerical solutions to the atomic Kohn-Sham equations as the secondary basis, and a scheme resembling geometry optimization to obtain the adaptive basis. The CONQUEST program [242] forms local "support functions" (an adaptive basis) from either functions akin to plane waves [243] or pseudo-atomic orbitals [244]. The ONETEP package [245, 246] forms non-orthogonal generalized Wannier functions (NGWFs) [247] as the environment-adapted basis, which is a linear combination of periodic sinc functions confined in an atom-centered sphere of fixed radius. The NGWFs are efficiently optimized via a preconditioned conjugate-gradient algorithm [248].

Recently, adaptive basis schemes that do not require the global Hamiltonian or density matrix have been presented. The localized filter diagonalization (LFD) method builds an adaptive basis on-the-fly by contracting the atomic Gaussian functions within a local region, with contraction coefficients determined by diagonalizing a block of the Hamiltonian matrix corresponding to that region [249, 250]. This algorithm has also been used to construct multisite local support functions [251], and the general philosophy has been extended by Lin et al. [252], including another model with more rigorous optimization [253]. While clearly promising, to our knowledge, the accuracy and performance of these methods on chemical systems have not been systematically assessed yet.

In the present work, we propose an inexpensive version of the PAO method (Sec. 2.2). Instead of energy-optimizing the adaptive basis and density simultaneously, an inexpensive converged SCF solution (density matrix) computed in a projected reference basis (PRB) is utilized as a reference (Sec. 2.2.1). Based on this reference, an atom-centered minimal adaptive basis (MAB) is found by minimizing a judiciously chosen surrogate function (Sec. 2.2.2), which only involves computationally inexpensive steps. The converged MAB is then used as the basis set for another SCF calculation, which requires small computational effort as well while providing comparable accuracy to PAO-SCF. Perturbation corrections (PC) can be applied to the MAB-SCF energy for obtaining the desired accuracy (Sec. 2.2.5). The overall MAB-SCF (PC) procedure is illustrated in Figure 2.2. Details about the pilot implementation of this scheme and proof-of-concept calculations are summarized in Sec. 2.3. As an

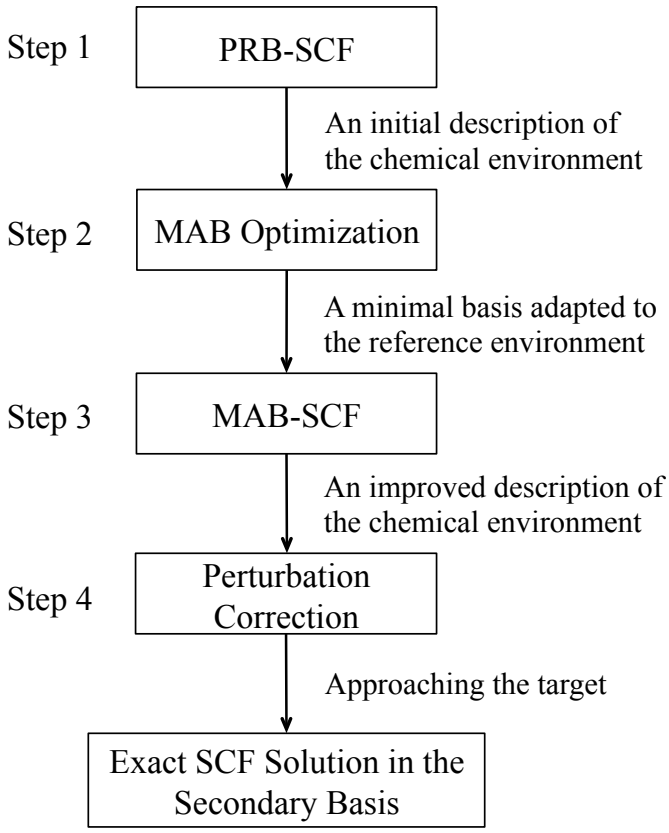

Figure 2.2: Illustration of the overall MAB-SCF (PC) procedure, which includes four steps in total.

approximate SCF method, its accuracy is assessed on a broad range of TC and NC datasets that is presented in Sec. 2.4 and 2.5.

2.2 Theory

The notation used throughout this paper is as follows: $|\omega\rangle$: generic atomic basis functions; $|\psi\rangle$: generic molecular orbitals; capital Roman indices X, Y,...: atomic centers; lowercase Greek letters μ, ν, λ,...: secondary (large) AO basis indices; α, β, γ,...: primary (PRB or MAB) AO basis indices; lowercase Romans i, j, k,...: occupied MO indices; a, b, c,...: virtual MO indices; p, q, r,...: generic MO indices. For introducing the MAB optimization scheme, i, j,... are also employed to denote the vectors retained in the MAB subspace, a, b,... for the vectors in MAB's complementary subspace, and p, q,... for the generic ones, which is analogous to the partitioning of MO space in SCF. The different basis sets that are involved in this work and the relationships between them are summarized in Table 2.1.

Unless otherwise specified, matrices in the secondary AO basis are denoted by bold Roman letters (e.g. $\mathbf{F}$, $\mathbf{P}$), while those in the primary basis are by bold calligraphic Roman letters (e.g. $\boldsymbol{\mathcal{F}}$, $\boldsymbol{\mathcal{P}}$). To concisely show the character of quantities within a nonorthogonal

Table 2.1: Summary of different level of atomic basis sets involved in the procedure of MAB-SCF.

	Name	Expression	Definition (Origin)
Standard Basis	Reference Basis (RB)	$\|\omega_{A\tilde{\alpha}}\rangle$	Standard double-ζ basis sets (e.g. 6-31+G(d))
	Secondary Basis (the target)	$\|\omega_{A\mu}\rangle$	Standard basis sets that are close to the CBS limit (e.g. aQZ)
Primary Basis	Projected Reference Basis (PRB)	$\|\omega_{A\alpha}\rangle = \|\omega_{A\mu}\rangle \left(B_{ref}\right)^{A\mu}_{\ A\alpha}$ (Eq. 2.1)	On-atom projection of RB into the secondary basis
	Minimal Adaptive Basis (MAB)	$\|\omega_{Ai}\rangle = \|\omega_{A\mu}\rangle B^{A\mu}_{\ Ai}$, $\mathbf{B} = \mathrm{argmin}\, E(\mathbf{B})$ *	Energetically optimized contraction of secondary basis functions on each atom

* $E(\mathbf{B})$ is the surrogate energetic objective function used for MAB optimization (defined by Eq. 2.10)

basis, tensorial notation will be used in the derivation, i.e., covariant (subscript) and contravariant (superscript) indices are distinguished, following Ref. 254 and the appendix of Ref. 236. For instance, a matrix element denoted by $B^{X\mu}_{\ X\alpha}$ indicates that matrix $\mathbf{B}$ has rows corresponding to contravariant secondary basis functions and columns corresponding to covariant primary basis functions, and these basis functions belong to the same atomic center X. Einstein summation convention is applied for contractions between contravariant and covariant indices, except for summations over different atomic centers, which will be written out explicitly.

2.2.1 SCF in the projected reference basis (PRB)

The search for the MAB described in Sec. 2.2.2 requires an inexpensively calculated reference density matrix in the secondary (target) basis. A converged SCF solution in a small PRB serves this purpose. The PRB is constructed by projecting the reference basis functions, $\{|\omega_{A\tilde{\alpha}}\rangle\}$, into the space spanned by the secondary basis on each atom[214]:

$$\begin{aligned} |\omega_{A\alpha}\rangle &= |\omega_{A\mu}\rangle \left(S_A^{-1}\right)^{A\mu A\nu} \left(S_{12}\right)_{A\nu A\alpha} \\ &= |\omega_{A\mu}\rangle \left(B_{ref}\right)^{A\mu}_{\ A\alpha}. \end{aligned} \tag{2.1}$$

Here, $\left(S_{12}\right)_{A\nu A\alpha} = \langle\omega_{A\nu}|\omega_{A\tilde{\alpha}}\rangle$ is the overlap between the (unprojected) reference basis and the secondary basis functions, while $\mathbf{S}_A$ is the overlap metric of secondary basis functions on atom A. Throughout this paper, $\mathbf{B}$ is used to denote atom-blocked matrices containing contraction coefficients of the secondary basis functions on each atomic site, which defines a primary basis.

Since the reference basis is small (to be specified later) while the secondary basis is close to the CBS limit, the block in $\mathbf{B}_{ref}$ will be very sparse since the contraction coefficients for the high angular momentum components of the secondary basis all vanish during the projection procedure. At this stage, an SCF calculation is performed in the PRB, by solving the following generalized eigenvalue equation:

$$\boldsymbol{\mathcal{F}}\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{S}}\boldsymbol{\mathcal{C}}\boldsymbol{\epsilon}, \tag{2.2}$$

$\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{S}}$ have the dimension of the PRB, and they can be transformed from their counterparts in the secondary basis using the $\mathbf{B}_{ref}$ matrix:

$$\boldsymbol{\mathcal{F}} = \mathbf{B}_{ref}^T \mathbf{F} \mathbf{B}_{ref}, \quad \boldsymbol{\mathcal{S}} = \mathbf{B}_{ref}^T \mathbf{S} \mathbf{B}_{ref} \tag{2.3}$$

In reverse, the PRB density matrix, $\boldsymbol{\mathcal{P}}$, can be projected into the secondary basis via the following transformation:

$$\mathbf{P}_{ref} = \mathbf{B}_{ref} \boldsymbol{\mathcal{P}} \mathbf{B}_{ref}^T \tag{2.4}$$

Since the PRB is an exact subset of the secondary basis, no information in $\boldsymbol{\mathcal{P}}$ is lost upon projection into the latter (Eq. 2.4), i.e., $\mathbf{P}_{ref}$ and $\boldsymbol{\mathcal{P}}$ contain the same information about the chemical environment (this is *not* true if $\boldsymbol{\mathcal{P}}$ is optimized in the reference basis directly without doing the projection). We call this special property of $\mathbf{P}_{ref}$ "PRB-representability". The final PRB density matrix $\boldsymbol{\mathcal{P}}$ therefore becomes the reference used in the search for the MAB.

2.2.2 Finding the minimal adaptive basis (MAB)

Sec. 2.2.1 employs a basis defined by a fixed atom-blocked transformation (the PRB) and converges a density matrix in it. With the fixed $\mathbf{P}_{ref}$ in hand, our goal now is to optimize an energy-like function with respect to a variable $\mathbf{B}$ matrix that defines the MAB. (Note: in the following discussion $\mathbf{B}$ *exclusively* denotes the MAB coefficients.) Since a single diagonalization minimizes Tr[$\mathbf{PF}$] for a chosen number of electrons [212] when $\mathbf{F}$ is given, we shall, by analogy, minimize Tr[$\tilde{\mathbf{P}}\mathbf{F}$], where $\tilde{\mathbf{P}}$ is a "MAB-representable" density matrix in the secondary basis:

$$\tilde{\mathbf{P}} = \mathbf{B} \boldsymbol{\mathcal{D}} \mathbf{B}^T. \tag{2.5}$$

$\boldsymbol{\mathcal{D}}$ is a density matrix in the MAB that derives from the fixed $\mathbf{P}_{ref}$.

However, the MAB has smaller rank than the PRB, and the spaces spanned by them are rather different, so $\mathbf{P}_{ref}$ will not be exactly representable by the MAB. There exist many possible ways to construct $\tilde{\mathbf{P}}$. We choose to project $\mathbf{P}_{ref}$ into the space spanned by the MAB first, which gives $\boldsymbol{\mathcal{D}}$, then transform it back into the secondary basis:

$$\boldsymbol{\mathcal{D}} = (\boldsymbol{\sigma}^{-1}) \mathbf{B}^T \mathbf{S} \mathbf{P}_{ref} \mathbf{S} \mathbf{B} (\boldsymbol{\sigma}^{-1}), \tag{2.6}$$

$$\tilde{\mathbf{P}} = \mathbf{B} (\boldsymbol{\sigma}^{-1}) \mathbf{B}^T \mathbf{S} \mathbf{P}_{ref} \mathbf{S} \mathbf{B} (\boldsymbol{\sigma}^{-1}) \mathbf{B}^T, \tag{2.7}$$

where $\boldsymbol{\sigma}$ refers to the overlap metric of the MAB

$$\boldsymbol{\sigma} = \mathbf{B}^T \mathbf{S} \mathbf{B}. \tag{2.8}$$

Recognizing that the projector into the MAB space is

$$\mathbf{R} = \mathbf{B} (\boldsymbol{\sigma}^{-1}) \mathbf{B}^T, \tag{2.9}$$

our surrogate energetic objective function becomes:

$$E = \mathrm{Tr} \left[\mathbf{R} \mathbf{S} \mathbf{P}_{ref} \mathbf{S} \mathbf{R} \mathbf{F} \right]. \tag{2.10}$$

For brevity, in the following derivation we use $\mathbf{P}$ instead of $\mathbf{P}_{ref}$ for the *fixed* reference density matrix.

We note that the MAB-representable density matrix $\tilde{\mathbf{P}}$ usually does not contain exactly the right electron count. While the exact N_{elec} is given by $\text{Tr}[\mathbf{PS}]$, utilizing the idempotency of $\mathbf{R}$ (based on Eq. 2.8 and 2.9, it is straightforward to derive $\mathbf{RSR} = \mathbf{R}$), we have

$$\begin{aligned} \text{Tr}[\tilde{\mathbf{P}}\mathbf{S}] &= \text{Tr}[\mathbf{RSPSRS}] \\ &= \text{Tr}[\mathbf{RSPS}] \neq \text{Tr}[\mathbf{PS}], \end{aligned} \tag{2.11}$$

The inequality arises because the reference density matrix $\mathbf{P}$ is usually not MAB-representable.

Gradient-based optimization can locate the optimal $\mathbf{B}$ as the minimizer of Eq. (2.10). The initial guess for the MAB (and its orthogonal complement in the span of the secondary basis functions on each atomic site, which is denoted by $\mathbf{V}$), is obtained by diagonalizing atomic blocks of the reference density matrix, appropriately transformed [223] as $\mathbf{P'_A} = \mathbf{X_A^T S_A P_A S_A X_A}$, where $\mathbf{S_A}$ is the overlap matrix of the secondary basis functions on A, and $\mathbf{X_A}$ is the canonical orthogonalizer for them. With $\mathbf{U_A}$ representing the eigenvectors of $\mathbf{P'_A}$, the initial $\mathbf{B}$ and $\mathbf{V}$ are set to:

$$(B_A^{\text{init}})^{A\mu}_{\ \ Ai} = (X_A)^{A\mu}_{\ \ Ap}(U_A)^{Ap}_{\ \ Ai} \tag{2.12a}$$

$$(V_A^{\text{init}})^{A\mu}_{\ \ Aa} = (X_A)^{A\mu}_{\ \ Ap}(U_A)^{Ap}_{\ \ Aa}, \tag{2.12b}$$

where $(U_A)^{Ap}_{\ \ Ai}$ denotes the eigenvectors corresponding to the m_A largest eigenvalues of $\mathbf{P'}_A$ (m_A is the rank of minimal basis for atom A), and $(U_A)^{Ap}_{\ \ Aa}$ denotes the remaining eigenvectors. This gives the initial partitioning of the Hilbert space that can be represented symbolically as

$$\begin{aligned} \mathbb{I}_{\mathbf{A}} &= \mathbf{X_A} = \mathbf{B_A} \oplus \mathbf{V_A}, \\ \mathbb{I} &= \bigoplus_{A=1}^{N_A} \mathbb{I}_{\mathbf{A}}. \end{aligned} \tag{2.13}$$

Since the MAB functions (and the complementary ones) are constructed by on-site contractions of the secondary basis, the variables that parameterize the MAB are intra-atomic orbital rotations. Akin to Ref. 200, a single on-block unitary transform is parameterized by the exponential of an antisymmetric matrix [13], which ensures that the updated atomic orbitals stay on the same manifold:

$$[B(\theta)]^{X\mu}_{\ \ Xi} = C^{X\mu}_{\ \ Xr} \exp\left(\theta^{Xr}_{\ \ Xi}\right), \tag{2.14}$$

where $C^{X\mu}_{\ \ Xr}$ denotes the union of the MAB and the complementary functions on atom X. To enforce antisymmetry of $\boldsymbol{\theta}$, it is further parameterized by $\boldsymbol{\Delta}$ which contains all the independent variables:

$$\begin{aligned} \theta^{Xr}_{\ \ Xi} &= \left[\Delta^{XrXs} - (\Delta^\dagger)^{XrXs}\right] \sigma_{XsXi} \\ &= \left[\Delta^{XrXs} - \Delta^{XsXr}\right] \sigma_{XsXi}. \end{aligned} \tag{2.15}$$

The desired gradient, evaluated at $\boldsymbol{\Delta} = \mathbf{0}$ is

$$\begin{aligned}\frac{\partial E}{\partial \Delta^{Z_pZ_q}} &= \sum_{X,Y} \frac{\partial E}{\partial R^{X_\mu Y_\nu}} \frac{\partial R^{X_\mu Y_\nu}}{\partial \Delta^{Z_pZ_q}} \\ &= \sum_{X,Y} (SPSRF + FRSPS)_{Y_\nu X_\mu} \frac{\partial R^{X_\mu Y_\nu}}{\partial \Delta^{Z_pZ_q}}. \end{aligned} \tag{2.16}$$

E is invariant with respect to orbital rotations within the MAB space ($p = i, q = j$), or within the space of complementary excluded vectors ($p = a, q = b$), as these rotations leave $\mathbf{R}$ unchanged. Therefore, the non-zero gradient comes only from variations of $\Delta^{Z_iZ_a}$. Using the identities

$$\frac{\partial B^{X_\mu}_{\ X_j}}{\partial \Delta^{Z_iZ_a}} = -V^{X_\mu}_{\ X_a} \delta^X_{\ Z} \sigma_{X_iX_j} \tag{2.17}$$

and

$$\frac{\partial (\boldsymbol{\sigma}^{-1})}{\partial \boldsymbol{\Delta}} = -(\boldsymbol{\sigma}^{-1}) \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\Delta}} (\boldsymbol{\sigma}^{-1}), \tag{2.18}$$

the desired gradient expression is given by

$$\frac{\partial E}{\partial \Delta^{Z_iZ_a}} = -2\sigma_{Z_iZ_j} \left[(\sigma^{-1}) B^T G (I - RS) \right]^{Z_j}_{\ Z_\mu} V^{Z_\mu}_{\ Z_a}, \tag{2.19}$$

where, for brevity, $\mathbf{G} = \partial E / \partial \mathbf{R}$ as defined in Eq. 2.16. More details about the derivation of Eq. (2.19) is provided in Appendix A.1.1.

Once the gradient at the current position is computed, the optimization algorithm will generate a new step ($\boldsymbol{\Delta}$) based on it (and the previous gradients and steps). The equations for the exponential transformation were derived in Ref. 13. The update for the MAB can be represented as

$$\mathbf{B}^{(n)} = \mathbf{B}^{(n-1)} \begin{pmatrix} \mathbf{U} \cos \boldsymbol{p}^{1/2} \mathbf{U}^{\mathbf{T}} \\ -\boldsymbol{\Delta}^{\dagger} \mathbf{U} \boldsymbol{p}^{1/2} \sin \boldsymbol{p}^{1/2} \mathbf{U}^{\mathbf{T}} \end{pmatrix}. \tag{2.20}$$

$\mathbf{U}$ and $\boldsymbol{p}$ stand for eigenvectors and eigenvalues of the matrix quantity $\boldsymbol{\Delta}\boldsymbol{\Delta}^{\dagger}$, respectively, and note that the unitary transformations are atom-blocked operations. When the iterative optimization converges, $\mathbf{B}$ represents a minimal basis energetically adapted to the chemical environment described by the reference density matrix (from PRB-SCF). Figure 2.3 illustrates the MAB optimization procedure. Finally, we note that for unrestricted cases, the MABs for α and β electrons are optimized separately (they are completely decoupled), using the same objective function form.

With the MAB defined, a converged SCF solution can be obtained in this basis. The SCF energy in the fixed MAB will be an approximation to the energy evaluated by PAO-SCF, which directly minimizes the SCF energy with respect to the generators of the MAB as well as the variables defining the density matrix. These two approaches will be compared in Sec. 2.4.

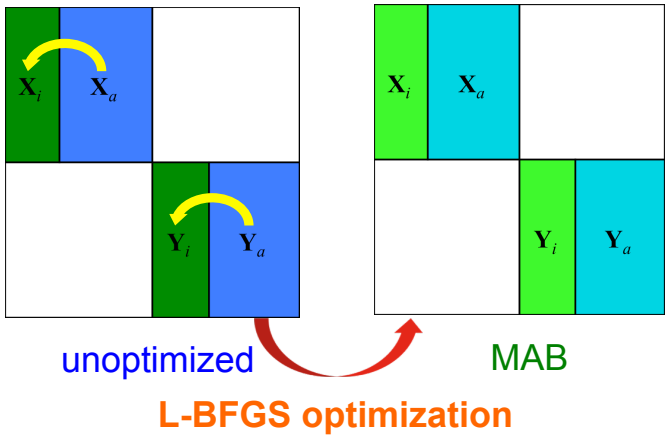

Figure 2.3: Illustration of the MAB optimization procedure. Atom centers X and Y correspond to two distinct diagonal blocks of the **B** matrix. X_i, Y_i represent the MAB functions on these sites, and X_a, Y_a represent the excluded basis functions.

2.2.3 Modified definition of the minimal adaptive basis

The size of a minimal basis only depends on the principal quantum number of the atom's valence shell, since a complete set of angular momentum functions are needed to fulfill the requirement of spatial isotropy. This definition usually works very well, but there are two types of exceptions. First, in some cases, the standard rank of the minimal basis includes redundant functions. For example, the minimal basis of lithium ($n = 2$) consists of 5 functions, although only two of them are required to describe its $1s^2 2s^1$ configuration. The same applies to many other electron-deficient species like cations and radicals. The presence of redundant functions causes difficulties in converging the MAB optimization procedure. Second, in some cases, the standard rank of the minimal basis is too small to accurately describe the bonding. Examples include some hypervalent molecules (e.g. SO_3, ClF_3, etc), and, occasionally, molecular anions. In such cases, the standard rank of the MAB will lead to larger errors in the resulting molecular energies, which can be greatly reduced if a certain number of additional MAB functions are judiciously added to the appropriate atomic centers.

In both cases, we can adjust the rank of the MAB appropriately based on information that is *already available* from the initial PRB-SCF calculation. The resulting procedure, shown in Algorithm 1, can either truncate or augment the MAB dimension on each atom. The number of significant eigenvalues (N_{Sig}) for each atom is set to the number of eigenvalues of $\mathbf{P}'_{\mathbf{A}}$ that are above a first threshold ($thresh1$, whose default value is 0.01). The MAB dimension will be reduced to N_{Sig} if that is smaller than a minimal basis (N_{min}). On the other hand, when $N_{min} < N_{Sig}$, the algorithm expands the MAB dimension by the number of eigenvalues beyond N_{min} that satisfy $E_A(i)/E_A(N_{min}) > thresh2$ ($thresh2$ has a default value of 0.02, i.e., eigenvalues that are larger than $E_A(N_{min})/50$ will be included), which will allow a lower optimized MAB-SCF energy. The default values of these two thresholds are

Algorithm 1: Algorithm that flexibly adjusts (truncates or expands) the size of the MAB for special systems, based on quantities already computed when generating the initial guess for the MAB optimization.

for *atom* $A = 1, N_A$ **do**
  Compute $\mathbf{P'_A} = \mathbf{X_A^T S_A P_A S_A X_A}$
  Diagonalize $\mathbf{P'_A}$, get its eigenvalues E_A
  $i = 1$, $N_{Sig} = 0$
  while $E_A(i) \geq thresh1$ **do**
    $i = i + 1$, $N_{Sig} = N_{Sig} + 1$
  $m_A = N_{min}^A$
  if $N_{Sig} < N_{min}^A$ **then**
    $m_A = N_{Sig}$ (truncated)
  else
    $i = N_{min}^A + 1$
    if $E_A(N_{min}^A) > 0.25$ **then**
      $thresh2$ /= 2
      (considering large absolute values)
    while $E_A(i)/E_A(N_{min}^A) \geq thresh2$ **do**
      $m_A = m_A + 1$ (augmented)
      $i = i + 1$
      if $i > dim(E_A)$ **then**
        break
  $\boldsymbol{m_A}$ becomes the rank of MAB on atom A

empirically determined based on the performance on the hypervalent molecules in the G2 set (see Sec. 2.4).

2.2.4 Modified MAB objective function

The objective function given by Eq. (2.10) can be rewritten as follows:

$$\begin{aligned} E &= \mathrm{Tr}\left[\mathbf{RSC}_o\mathbf{C}_o^T\mathbf{SRF}\right] \\ &= \mathrm{Tr}\left[\mathbf{C}_o^T\mathbf{SRFRSC}_o\right] \\ &= \mathrm{Tr}[\tilde{\mathbf{C}}_o^T\mathbf{F}\tilde{\mathbf{C}}_o], \end{aligned} \tag{2.21}$$

where $\tilde{\mathbf{C}}_o = \mathbf{RSC}_o$ represents the PRB-optimized occupied MOs after being projected into the MAB space. For stable species, the energies of occupied MOs should all be negative, and thus minimizing E corresponds to retaining as many of the bound electrons as possible.

A disaster occurs in the MAB optimization if an occupied MO has a positive energy, because minimization will result in loss of those electrons. With inexact functionals, this occasionally happens for anions. For example, the energy of the three $2p$ orbitals in F^- is 0.001 Eh with the B3LYP functional (hybrid), and 0.056 Eh with B97-D (pure). With these functionals, the resulting value of Tr[**RSPS**] (number of electrons captured by $\tilde{\mathbf{P}}$) is close to 2 when the MAB is optimized, which indicates that the six $2p$ electrons are missing! This is completely unphysical, and causes the SCF energy computed with the MAB to be qualitatively incorrect.

Such difficulties can be avoided by modifying the eigenvalue structure of $\mathbf{F}$ to ensure that all occupied levels are negative. This can be done by applying a uniform shift to all the eigenvalues:

$$\mathbf{F}' = \mathbf{F} - \lambda\mathbf{S} \tag{2.22}$$

using $\mathbf{C^T SC} = \mathbf{I}$. The shift, λ, is set to be:

$$\lambda = \varepsilon(\text{HOMO}) + \alpha\left[\varepsilon(\text{LUMO}) - \varepsilon(\text{HOMO})\right], \tag{2.23}$$

where $0 < \alpha \leq 1$ so that the zero energy lies between the HOMO and LUMO calculated by PRB-SCF (note: such a shift is applied *only* when ϵ(HOMO) is detected to be positive). The "mixing" parameter $\alpha = 0.75$ is empirically selected based on the performance of our method on small (monoatomic and diatomic) anions (see Appendix A.2.1 for more details). Replacing $\mathbf{F}$ with $\mathbf{F}'$ in Eq. (2.10) gives a modified objective function for the MAB:

$$\begin{aligned} E &= \text{Tr}\left[\mathbf{RSPSR}(\mathbf{F} - \lambda\mathbf{S})\right] \\ &= \text{Tr}[\mathbf{RSPSRF}] - \lambda\,\text{Tr}[\mathbf{RSPS}]. \end{aligned} \tag{2.24}$$

When $\lambda > 0$, the new term resembles a penalty for losing electrons, which can be made explicit by adding an additional constant, λN_{elec}, to the RHS:

$$E = \text{Tr}[\mathbf{RSPSRF}] + \lambda\left(N_{\text{elec}} - \text{Tr}[\mathbf{RSPS}]\right). \tag{2.25}$$

2.2.5 Perturbation correction schemes

Based on the data presented in Ref. 236 and 237, a significant difference exists between PAO-SCF and exact SCF results. To reduce this gap, computationally inexpensive correction schemes based on perturbation theory are useful. Analogous to the dual-basis methods, the converged MAB-SCF solution serves as the primary basis reference, and the contribution of non-Brillouin singles to the second-order perturbative (PT2) energy correction is given by [211]

$$\delta E = \text{Tr}\left[\mathbf{F}^{(1)}_{OV}\mathbf{T}^{(1)}_{VO}\right]. \tag{2.26}$$

Here $\mathbf{F}$ denotes the Fock matrix built upon the MAB-SCF density projected into the secondary basis: $\mathbf{F} = \mathbf{F}(\tilde{\mathbf{P}})$. The first-order T-amplitude satisfies the following linear equation:

$$\mathbf{F}^{(1)}_{VO} + \mathbf{F}^{(0)}_{VV}\mathbf{T}^{(1)}_{VO} - \mathbf{T}^{(1)}_{VO}\mathbf{F}^{(0)}_{OO} = \mathbf{0}^{(1)}_{VO}. \tag{2.27}$$

In the pseudo-canonicalized MO basis (obtained by diagonalizing $\mathbf{F}_{OO}$ and $\mathbf{F}_{VV}$ separately, see Appendix A.1.3), Eq. (2.27) reduces to a simpler form:

$$T_{ai}^{(1)} = -F_{ai}^{(1)} / \left(\epsilon_a^{(0)} - \epsilon_i^{(0)}\right). \tag{2.28}$$

Correspondingly, the perturbative energy lowering becomes

$$\delta E = -\sum_{ia} |F_{ai}^{(1)}|^2 / \left(\epsilon_a^{(0)} - \epsilon_i^{(0)}\right), \tag{2.29}$$

which can be interpreted as an energy-weighted steepest descent (or an approximate Newton) step [211, 237]. Alternatively, other corrections that involve a full diagonalization of the Fock matrix can be applied, such as the aforementioned DB-SCF (only slightly different from PT2) and DFPC methods. The latter performs a single update of the density matrix in the secondary basis (by diagonalizing $\mathbf{F}$), and then recomputes the full SCF energy based upon that density matrix (the result will thus be variational).

2.3 Computational Details

A pilot implementation of this new SCF scheme is accomplished in a development version of the Q-Chem 4.3 package [255]. A preconditioned limited-memory BFGS (L-BFGS) algorithm [256, 257] is implemented for solving the MAB optimization problem efficiently. The inverted on-diagonal blocks of the Hessian matrix for the objective function (second derivatives with regard to orbital rotations on the same atomic site) are employed as the preconditioner of the L-BFGS algorithm. In most scenarios, this preconditioning strategy leads to convergence of the MAB optimization in a reasonable number of iterations ($10^1 \sim 10^2$), with only moderate additional cost to evaluate the preconditioner. More details about the preconditioned L-BFGS algorithm and the evaluation of the on-diagonal blocks of the Hessian are provided in Appendix A.1.2.

In the current implementation, all the density matrix updates are computed by diagonalizing $\boldsymbol{\mathcal{F}}$ (Fock matrix in the dimension of PRB or MAB), and the only diagonalization in the secondary basis dimension ($N_v \times N_v$ to be exact) is required by the perturbation correction. However, to obtain $\boldsymbol{\mathcal{F}}$ (for the time being) still requires contracting the ERI tensor with a PRB- or MAB-representable density matrix in the secondary basis (using restricted KS Fock matrix as an example):

$$F_{\mu\nu} = h_{\mu\nu} + 2\,(\mu\nu|\lambda\sigma)\,\tilde{P}^{\lambda\sigma} - \kappa\,(\mu\sigma|\lambda\nu)\,\tilde{P}^{\lambda\sigma} + (V_{xc})_{\mu\nu}, \tag{2.30}$$

where κ is the proportion of exact exchange in the employed functional. Then $\mathbf{F}$ is transformed back into the primary basis through Eq. 2.3. This choice is actually less efficient, because quantities in the primary basis, like $\mathcal{P}^{\alpha\beta}$, $\mathcal{C}_i^{\alpha}$, can be directly utilized to construct $\boldsymbol{\mathcal{F}}$, which will significantly reduce the dimension of the contraction. Therefore, with our preliminary implementation, we will focus on validating the accuracy of MAB-SCF (PC) in

this work, and the potential efficient implementation of this method will be briefly discussed in Sec. 2.6.

All the results for TC and NC datasets are generated with the KS-DFT routines in Q-Chem 4.3 as well. A $(75, 302)$ grid (75 radial shells with 302 Lebedev points in each) is used for all employed exchange-correlation (XC) functionals, and the SG-1 grid [258] is used for the VV10 [71] NLC functional. Unless otherwise noted, 6-31+G(d) is used as the reference basis in PRB-SCF. Smaller reference bases could alternatively be employed, at the cost of diminished accuracy (see the results provided in Appendix A.2.2). The optimization of the MAB converges to 10^{-6} a.u., while all the SCF calculations converge to 10^{-8} a.u. (RMS of the gradient). To determine the appropriate dimension of the MAB, the default values of $thresh1$ and $thresh2$ adopted by Algorithm 1 are set to 0.01 and 0.02, respectively. We note that in this work, the "adding vector" strategy is by default turned *off* and only utilized for specified hypervalent molecules.

2.4 Preliminary Tests on G2 Thermochemistry

2.4.1 Comparison with PAO-SCF

We start investigating the accuracy of our method by performing a series of preliminary tests on the G2 thermochemistry set [185]. To test the quality of the optimized MAB, the performance of MAB-SCF on the G2 set is compared to PAO-SCF, since the latter gives the limiting behavior of an atom-centered minimal basis. For the reasons discussed in Sec. 2.2.3, the minimal basis models (including MAB and PAO) are not sufficient for describing hypervalent molecules. Thus, we designate molecules containing Al, Si, P, S, Cl centers that are coordinated by highly electronegative atoms (e.g. O, F, Cl) as hypervalent, including SO, ClO, SO_2, AlF_3, $AlCl_3$, SiF_4, $SiCl_4$, PF_3, ClF_3, and $(CH_3)_2SO$, and exclude them from the test set preliminarily.

The results for this "pruned" G2 set (138 molecules) computed with three functionals are collected in Table 2.2 (aQZ is the target secondary basis). To make it a fair comparison, molecules that fail to converge their PAOs (listed in the table footnotes) are also excluded when reporting the statistical errors. At the SCF level (i.e. without PT2 correction), PAO-SCF and MAB-SCF show similar accuracy with respect to the exact SCF results for all three tested functionals. Surprisingly, the MAB-SCF results are slightly better as a result of error cancellation (PAO-SCF is exact for atomic energies). Applying the PT2 correction significantly reduces the errors of both schemes. For the two pure functionals (B97-D and B97M-V), the RMSDs of MAB+PT2 are smaller than 0.1 kcal/mol ($\sim$ 0.05 kcal/mol), and they are close to those of PAO+PT2. For B3LYP, the RMSDs of both schemes noticeably increase, which suggests diminished effectiveness of PT2 when hybrid functionals are used. Nevertheless, we notice that the performance of MAB+PT2 is rather similar to that of PAO+PT2.

To better compute the energies of the hypervalent molecules, we increase the dimensions

Table 2.2: Summary of the errors of PAO-SCF and MAB-SCF (without and with PT2 correction) for the "pruned" G2 set against the conventional SCF results. aQZ is employed as the secondary (target) basis. Maximum errors (MAX), root-mean-square deviations (RMSD), and mean-signed errors (MSE) are reported in kcal/mol. Molecules where PAO failed to converge (listed in the footnotes) are excluded for both PAO-SCF and MAB-SCF when evaluating the statistical errors.

	B97-D[a]		B97M-V[b]		B3LYP[c]	
SCF energies						
	MAB	PAO	MAB	PAO	MAB	PAO
MAX	25.11	25.48	26.70	26.96	24.39	24.85
RMSD	7.07	7.27	8.06	8.07	7.05	7.18
MSE	5.59	5.72	6.54	6.49	5.62	5.66
With PT2 correction						
	MAB	PAO	MAB	PAO	MAB	PAO
MAX	-0.26	-0.14	-0.14	-0.08	0.84	0.91
RMSD	0.06	0.03	0.03	0.02	0.18	0.20
MSE	-0.02	0.00	-0.02	-0.01	0.14	0.15

[a] Convergence failures: ·CCH
[b] Convergence failures: SO_2, ClF_3, ·SH
[c] Convergence failures: NaCl

of their MAB based on Algorithm 1. The modified MAB+PT2 results are compared with those using the standard minimal basis dimensions in Table 2.3. For a majority of these molecules (ClO, SiF_4, $SiCl_4$, PF_3, $(CH_3)_2SO$, and presumably SO_2 and ClF_3), the errors are reduced by over 10 times by using the "adding vector" strategy. The degree of inadequacy of the conventional minimal basis dimension is perhaps a measure of molecular hypervalency. Indeed, species like SiF_4, PF_3 do not formally violate the "octet" rule, which indicates that molecular hypervalency may exist beyond its usual definition. On the other hand, $AlCl_3$ and AlF_3 do not seem to be typical hypervalent species, because use of standard minimal basis dimensions does not result in errors as large as those for the other molecules listed in Table 2.3.

Combining these specially treated hypervalent molecules with the other 138 molecules computed with the standard MAB model, the overall RMSD of MAB+PT2 against conventional SCF results for the G2 set is 0.033 kcal/mol. This result is only minimally different from the RMSD for the "pruned" G2 set (with B97M-V), by contrast with the poor results for standard MAB+PT2 and PAO+PT2 when hypervalent molecules are included. Therefore in the later G2 tests, unless otherwise specified, statistical errors evaluated including *all* the molecules will be reported with the hypervalent ones *separately treated* via Algorithm 1.

Table 2.4 shows how the dimensions of the MAB are increased on the central atoms of the hypervalent molecules after applying Algorithm 1. According to the rightmost column, the

Table 2.3: Errors (in kcal/mol) for the atomization energies of the hypervalent molecules in the G2 set computed with B97M-V. aQZ is used as the secondary basis. Results of unmodified MAB+PT2, MAB+PT2 with the "adding vector" strategy, and PAO+PT2 are compared against exact SCF results. The corresponding statistical errors for the full G2 set (including these molecules) are also reported.

	MAB+PT2 (normal)	MAB+PT2 (add_vec)	PAO+PT2
SO	0.049	0.022	0.085
ClO	-0.182	0.010	-0.120
SO_2	-0.990	0.001	N/A
AlF_3	-0.030	0.002	0.002
$AlCl_3$	0.050	-0.016	-0.017
SiF_4	-0.214	-0.007	-0.268
$SiCl_4$	-0.445	-0.009	-0.544
PF_3	-0.720	0.009	-0.544
ClF_3	-1.857	-0.020	N/A
$SO(CH_3)_2$	-1.173	-0.026	-1.061
G2 statistics (all molecules)			
MAX	-1.857	-0.141	-1.061
RMSD	0.238	0.033	0.114
MSE	-0.058	-0.017	-0.024

modified MAB function counts are usually close to those of a minimal basis plus one set of d (polarization) functions. Although AlF_3 and $AlCl_3$ do not show strong hypervalent character according to Table 2.3, the Al atom nonetheless gains additional MAB basis functions.

In general, the MAB optimization problem is considerably easier to converge than the aforementioned "double" PAO-SCF optimization. In contrast to PAO-SCF that encounters several convergence problems (mentioned in Table 2.2), no MAB convergence failure is detected for the entire G2 set with all three tested functionals. Furthermore, in contrast to PAO-SCF, the MAB optimization is decoupled from density matrix optimization, and thus MAB iteration counts will not directly affect the required number of SCF cycles. This is extremely important because of the much more significant cost per iteration for the latter. Table 2.5 lists molecules in the G2 set that require over 100 iterations to converge their PAOs (using B97M-V/aQZ). Apart from several convergence failures, the PAO-SCF scheme requires an enormous number of Fock matrix constructions for some molecules, such as NaCl. The MAB scheme, on the other hand, attains the optimized adaptive basis in fewer iterations for most of these molecules. Even in cases like COS and C_2H_4O (oxirane) where the iteration counts for optimizing the adaptive basis are similar, MAB-SCF is still far more efficient because many fewer Fock builds are required. Therefore, MAB-SCF appears to be

Table 2.4: Modifications of the MAB dimensions on the central atoms of the hypervalent molecules in the G2 set after the "adding vector" strategy is applied. The number of MAB functions for the coordinating atoms also occasionally changes (not shown).

Molecule	Central atom	*dim* (MAB) (original)	*dim* (MAB) (add_vec)
AlF_3	Al	9	16
$AlCl_3$	Al	9	13
SiF_4	Si	9	15
$SiCl_4$	Si	9	15
PF_3	P	9	14
SO	S	9	13
SO_2	S	9	13
$(CH_3)_2SO$	S	9	12
ClO	Cl	9	12
ClF_3	Cl	9	14

a more feasible adaptive basis SCF scheme than PAO-SCF, with comparable accuracy.

2.4.2 Functional dependence: pure vs hybrids

Table 2.2 already suggests that the performance of MAB+PT2 is not completely functional-independent. A clear difference exists between using pure and hybrid functionals. Therefore, we must investigate the performance of this method when different flavors of density functionals are employed. Using the G2 set with aug-cc-pVQZ as the secondary basis, Table 2.6 explores the performance of 13 density functionals. The first seven functionals do not contain exact exchange, including three GGAs (B97-D, BLYP[34, 35], PBE [36]) and four meta-GGAs (TPSS[40], MGGA_MS1 [260], M06-L [41], B97M-V). For these functionals, the MAB+PT2 scheme demonstrates good accuracy (M06-L is the largest outlier), while the RMSDs computed by MAB+DFPC are roughly twice as large. Thus PT2 appears preferable.

The other six functionals in Table 2.6 are hybrid functionals, including TPSSh (10%) [261], B3LYP (20%), PBE0 (25%) [43], M06-2X (54%), ωB97X-D (RSH) [49], and ωB97X-V (RSH) ("%" denotes the proportion of exact exchange). They show results that contrast with those of the pure functionals. For these hybrids, PT2 undershoots the exact SCF energy, and the size of the errors grows roughly with the amount of exact exchange. This leads to unsatisfactory accuracy (RMSD $>$ 0.2 kcal/mol) for functionals that contain more exact exchange than B3LYP. The MAB+DFPC approach, on the other hand, shows comparatively better performance across the hybrids (RMSDs are around 0.1 kcal/mol), than PT2. For the two RSH functionals, the results of MAB+DFPC are 5–6 times more accurate than

Table 2.5: Molecules in the G2 set that require over 100 iterations to converge their PAOs. The counts of iterative optimization steps and Fock builds required to converge PAO-SCF and MAB-SCF are compared. For several radicals (CH, ·SH, CH_3CH_2O·), considerably larger number of iterations are required by the PRB-SCF step because the geometric direct minimization (GDM) algorithm [259] is utilized to circumvent unstable SCF solutions.

	Num of opt steps		Num of Fock builds		
Molecules	PAO	MAB	PAO	PRB	MAB
CH	277	93	277	27	6
Na_2	1219	23	1219	7	3
Si_2	145	20	145	16	22
NaCl	3208	47	3208	9	5
SO_2	N/A	212	N/A	11	8
COS	148	146	148	10	7
ClF_3	N/A	83	N/A	12	8
C_2Cl_4	614	80	614	9	5
C_4H_6 (2-butyne)	490	149	490	10	7
C_2H_4O	109	111	109	9	6
SH	N/A	60	N/A	48	8
CH_3CH_2O	450	156	450	44	9

Table 2.6: RMSDs (in kcal/mol) of MAB-SCF with two different perturbation correction schemes (PT2 and DFPC) for the G2 set. Different pure (B97-D, BLYP, PBE, TPSS, MGGA_MS1, M06-L, B97M-V) and hybrid (TPSSh, B3LYP, PBE0, M06-2X, ωB97X-D, ωB97X-V) functionals are investigated. aQZ is employed as the secondary basis for all the calculations.

Functionals	MAB+PT2	MAB+DFPC
B97-D	0.053	0.115
BLYP	0.074	0.148
PBE	0.065	0.132
TPSS	0.055	0.119
MGGA_MS1	0.042	0.108
M06-L	0.144	0.156
B97M-V	0.033	0.079
TPSSh	0.099	0.086
B3LYP	0.181	0.083
PBE0	0.220	0.066
M06-2X	0.458	0.076
ωB97X-D	0.645	0.119
ωB97X-V	0.668	0.133

MAB+PT2.

We conclude that DFPC should be used as the correction to MAB-SCF when hybrid functionals are employed, at the expense of one more Fock build in the secondary basis. On the other hand, MAB+PT2 is less expensive and more accurate for pure functionals.

2.4.3 Basis set convergence

Different basis sets are employed for KS-DFT calculations, based on considerations such as accuracy, efficiency, and user experience. With the B97M-V functional, we assess the performance of MAB+PT2 with several widely used basis sets, including those in Dunning's correlation-consistent series (aTZ, QZ, aQZ)[178, 179], Jensen's polarization-consistent series (apc-2, pc-3, apc-3)[262–264], and the Karlsruhe "def2" series (TZVPPD, QZVPP, QZVPPD)[183]. The popular "large Pople" basis set 6-311++G(3df,3pd) [265, 266] is also included. These basis sets are augmented triple-ζ quality or higher, because our goal is to approach CBS limit results.

The RMSDs for the G2 set using different target secondary basis sets are displayed in Figure 2.4. The errors are typically below 0.1 kcal/mol, which indicates excellent transferability. The best performance is achieved by two quadruple-ζ basis sets with diffuse functions, aQZ and QZVPPD. For their unaugmented counterparts, QZ and QZVPP, the RMSDs are slightly larger, although still satisfactory. Since diffuse functions typically have no major impact on the accuracy of evaluated energetics for bonded interactions, the compatibility of MAB+PT2 with these unaugmented basis sets is helpful. The largest RMS errors are produced by Jensen's pc-3 and aug-pc-3 basis sets. This is mostly due to the inaccurate atomic energy for Li: the RMSDs are reduced to 0.063 (pc-3) and 0.059 (apc-3) kcal/mol if we exclude three Li-containing molecules (Li_2, LiF, and LiH) from the G2 set. As these are high-quality basis sets, the outlier might be due to poor compatibility of pc-3/apc-3 with the employed reference basis, 6-31+G(d), for the Li atom. Computational cost aside, any of the largest basis sets approach the true CBS limit and also the top accuracy of our method.

Ultimately, MAB-SCF with a perturbation correction scheme (MAB-SCF (PC)) might replace conventional SCF in KS-DFT calculations to approach the CBS limit. Therefore, we want to see how MAB-SCF (PC) approaches the CBS limit with the increasing size of the secondary basis. For that purpose, we extend Figure 2.5 (which motivates this work) with the functional RMSDs evaluated by MAB+PT2 (for B97-D, B97M-V) and MAB+DFPC (for B3LYP, M06-2X, ωB97X-V), as shown in Figure 2.5. These results are quite encouraging: the conventional SCF convergence towards the CBS limit of each functional is closely reproduced by MAB-SCF (PC). At the aQZ level, the largest difference between MAB-SCF (PC) and exact SCF results is only about 0.03 kcal/mol (for B97-D and B3LYP), which is below 1% of the intrinsic (CBS) error of the functional itself, and thus negligible. Larger differences between these two sets of results exist at the aDZ level (the largest gap is 0.26 kcal/mol for ωB97X-V/aDZ), but this is not relevant to our target of the CBS limit.

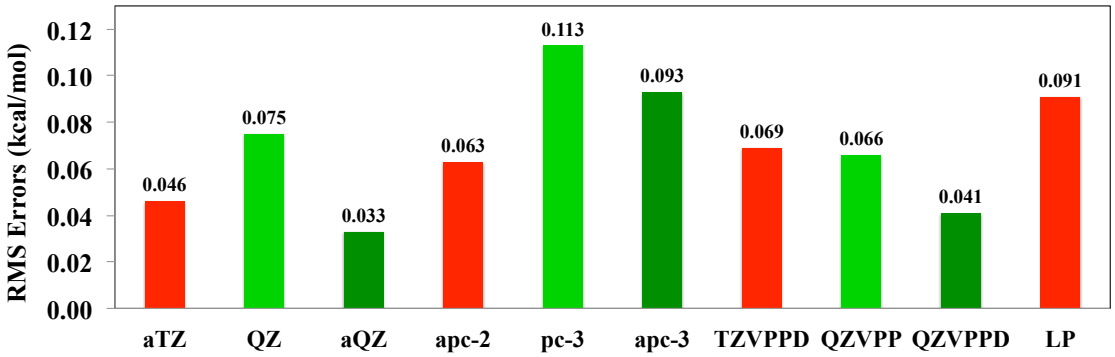

Figure 2.4: Assessment of the performance of MAB+PT2 with 10 secondary basis sets. RMSDs (vs. exact SCF) for the G2 set calculated with the B97M-V functional are reported. Different colors are used to indicate distinct basis set categories: augmented triple-ζ (red), unaugmented quadruple-ζ (light green), and augmented quadruple-ζ (dark green).

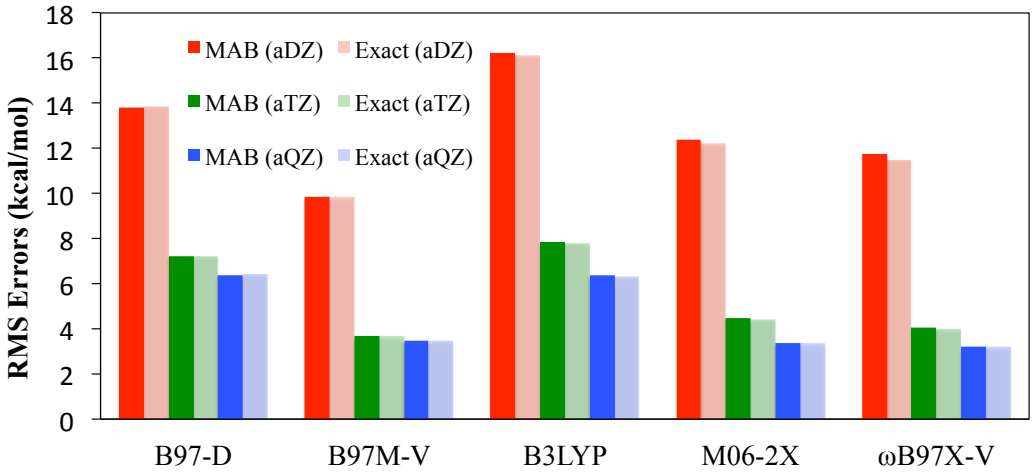

Figure 2.5: Basis set convergence of the functional RMSDs for the G2 set evaluated by MAB-based methods (solid) and conventional SCF (transparent). The reference values are taken from Ref. 186. Dunning's augmented correlation-consistent basis sets are utilized to approach the CBS limit (D: red, T: green, Q: blue). PT2 and DFPC correction schemes are applied on top of the MAB-SCF results of local (B97-D, B97M-V) and hybrid (B3LYP, M06-2X, ωB97X-V) functionals, respectively.

Table 2.7: RMSDs (in kcal/mol) of MAB-SCF (PC) for the W4-11 dataset (vs. exact SCF) evaluated with using three different functionals. PT2 is applied for B97-D and B97M-V while DFPC is used for B3LYP. aQZ is used as the secondary basis.

	B97-D	B97M-V	B3LYP
BDE99	0.061	0.038	0.070
HAT707	0.101	0.061	0.100
ISO20	0.068	0.048	0.085
SN13	0.034	0.018	0.053
TAE140	0.071	0.046	0.096
Overall	0.093	0.057	0.096

2.5 Additional Accuracy Tests

2.5.1 Thermochemistry

We will assess how the performance of MAB-SCF (PC) transfers to other thermochemistry (TC) datasets. Three density functionals (B97-D, B97M-V, B3LYP) will be employed to examine the accuracy of MAB-SCF (PC), using aQZ as the secondary basis, and the perturbation correction schemes are applied in the same way as in Figure 2.5. First we consider the W4-11 dataset [267], which includes 99 bond dissociation energies (BDE99), 707 heavy-atom transfer reaction energies (HAT707), 20 isomerization energies (ISO20), 13 nucleophilic substitution reaction energies (SN13), and 140 total atomization energies (TAE140). Note that the multi-reference (MR) species in W4-11 are included in this test. Thirteen species are separately treated as hypervalent molecules: AlF_3, $AlCl_3$, SiF_4, P_4, SO, SO_2, SO_3, S_2O, S_2, S_3, S_4, ClO, and OClO (see the previous discussion for the G2 set).

Table 2.7 contains the RMSDs of the MAB-SCF (PC) approach (against exact SCF) for W4-11, where different TC categories have been separated. The overall performance is similar to that for the G2 set, which shows encouraging transferability. Taking B97M-V as an example, the smallest and largest RMSDs of MAB+PT2 (vs. exact SCF) are obtained on SN13 (0.02 kcal/mol) and HAT707 (0.06 kcal/mol), respectively, while the corresponding functional RMSDs (vs. W4 reference) are 1.39 kcal/mol and 3.90 kcal/mol. Therefore, the errors caused by replacing conventional SCF with MAB-SCF (PC) are usually smaller than intrinsic functional errors by one or two orders of magnitude. A more straightforward comparison is provided by Figure 2.6: for B97M-V, the functional RMSDs computed via the MAB+PT2 approach show almost no difference compared to those by normal SCF method.

We also examined the performance of MAB-SCF (PC) on other TC datasets, including adiabatic ionization potentials and electron affinities (21 for each: G21IP and G21EA)[268, 269], 38 non-hydrogen transfer and 38 hydrogen transfer barrier heights (NHTBH38 [270] and HTBH38 [271]), and 14 alkane isomerization energies (Pentane14 [272]). The computational details are identical to those for W4-11, except for the anions in G21EA, HTBH38

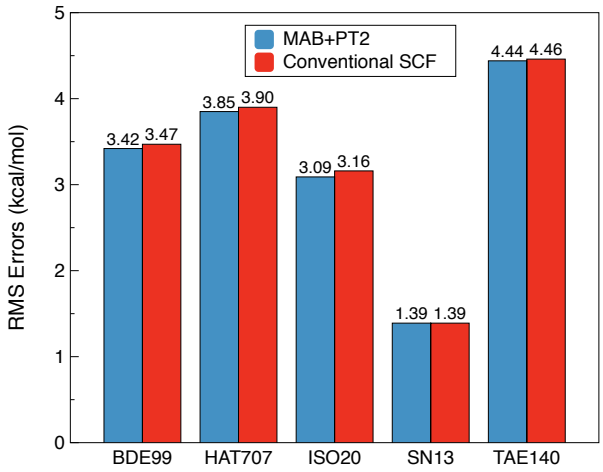

Figure 2.6: B97M-V's RMSDs for W4-11 (vs. reference values) evaluated by MAB+PT2 (blue) and conventional SCF (red). Very similar accuracy is obtained by these two SCF schemes at the B97M-V/aQZ level of theory.

and NHTBH38, where the modified MAB objective function (introduced in Sec. 2.2.4) is automatically applied to avoid the emergence of unphysical results. The RMSDs of MAB-SCF (PC) are presented in Table 2.8. We see that HTBH38 and Pentane14 are relatively easier cases for MAB-SCF (PC) to approximate the exact SCF result; and for G21IP and NHTBH38, the size of the RMSDs is similar to that for those W4-11 subsets (e.g. BDE99, SN13, etc.).

The largest RMSD appears on the G21EA dataset. Although none of the results are qualitatively incorrect by applying the modified MAB objective function, there are several molecular anions whose absolute energies evaluated by MAB-SCF (PC) are rather unsatisfactory: NO^-, PO^-, O_2^-, and S_2^-. Based on the discussions in Sec. 2.2.3, the "adding vector" strategy may also be applied for these electron-abundant species. As a result, the absolute energies of these anions are significantly improved, as shown in Table. 2.9. With these 4 molecular anions specially treated, the RMSDs for the G21EA dataset are recalculated and the results are also presented in Table 2.8, which turn out to be more comparable to the RMSDs for other TC test sets.

2.5.2 Non-covalent interactions

One of the key improvements in modern density functionals is for non-covalent interactions (NC). Therefore, we assess the performance of MAB-SCF (PC) on several NC datasets, including A24 (24 small NC complexes) [180], S22 (22 diverse small- to medium-sized NC complexes at the equilibrium geometries) [273, 274], HB15 (15 ionic hydrogen bond interactions) [275], H2O6Bind8 (binding energies of eight configurations of water hexamers) [76, 276], and FmH2O10 (binding energies of 10 configurations of $F^-(H_2O)_{10}$ [76, 276]). Since

Table 2.8: RMSDs (in kcal/mol) of MAB-SCF (PC) for TC datasets other than W4-11: G21IP, G21EA, HTBH38, NHTBH38, Pentane14. The numbers in parentheses are recalculated RMSDs for G21EA after applying the "adding vector" strategy for four diatomic anions: NO^-, PO^-, O_2^-, and S_2^-.

	B97-D	B97M-V	B3LYP
G21IP	0.050	0.036	0.040
G21EA	0.598 (0.101)	0.412 (0.027)	0.455 (0.078)
HTBH38	0.016	0.035	0.031
NHTBH38	0.099	0.071	0.118
Pentane14	0.005	0.007	0.002

Table 2.9: Errors in kcal/mol (vs. exact SCF) for the absolute energies of four diatomic anions (NO^-, PO^-, O_2^-, and S_2^-) evaluated with standard MAB-SCF (PC) ("Standard") and with the "adding vector" strategy turned on ("Add_Vec"). Other computational details are the same as in Table 2.8.

	Standard			Add_Vec		
	B97-D	B97M-V	B3LYP	B97-D	B97M-V	B3LYP
NO^-	-2.36	-1.63	1.60	-0.13	-0.06	0.10
PO^-	-0.52	-0.55	0.92	-0.02	-0.02	0.17
O_2^-	-2.04	-1.25	1.66	-0.01	-0.02	0.07
S_2^-	-0.41	-0.16	0.23	-0.01	0.02	0.02

many NC interactions have smaller magnitudes than TC energy differences, and some modern density functionals are able to achieve very small errors for them (e.g. B97M-V's unweighted RMSD for 1458 non-covalent interactions is 0.22 kcal/mol [73]), higher accuracy is needed for the MAB approach to match the exact SCF results.

As before, three density functionals (B97-D, B97M-V, B3LYP-D3(0)) are employed to assess the performance of MAB-SCF (PC) on these NC datasets. To avoid counterpoise (CP) corrections, we choose def2-QZVPPD as the secondary basis, which has fewer functions than aQZ but generates even smaller BSSEs [87]. Table 2.10 contains the resulting RMSDs (vs. exact SCF) for these NC datasets. Very small MAB-SCF (PC) errors are found for dimer binding energies (A24, S22, and HB15), with all three functionals. This is very encouraging for treating the most common non-covalent interactions by MAB-SCF (PC) instead of conventional SCF. Larger errors appear for the cluster binding energies (H2O6Bind8, FmH2O10), due to the larger magnitude of these interactions (H2O6Bind8: -40 to -50 kcal/mol; FmH2O10: about -170 kcal/mol) and the uniformity of their interaction types

Table 2.10: RMSDs (in kcal/mol) of MAB-SCF (PC) for five NC datasets: A24, S22, HB15, H2O6Bind8, and FmH2O10. Data points computed by conventional SCF using the def2-QZVPPD basis set (noCP) provide the reference values.

	B97-D	B97M-V	B3LYP-D3(0)
A24	0.009	0.005	0.009
S22	0.015	0.023	0.025
HB15	0.021	0.027	0.023
H2O6Bind8	0.061	0.085	0.118
FmH2O10	0.078	0.026	0.350

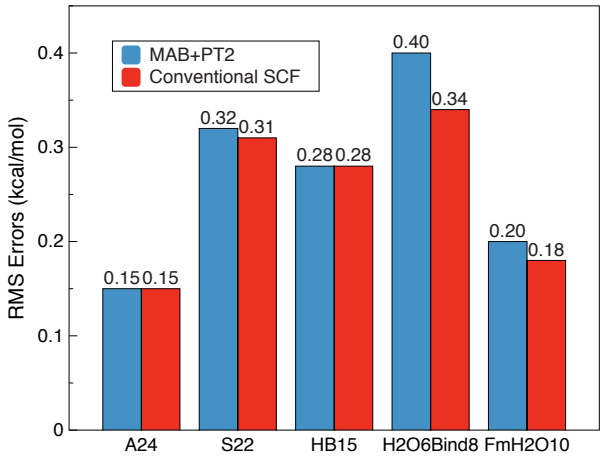

Figure 2.7: B97M-V's RMSDs for 5 NC datasets (vs. reference values) evaluated by MAB-SCF (PC) (blue) and conventional SCF (red), respectively. def2-QZVPPD is employed as the secondary basis set without using counterpoise corrections.

(systematic errors accumulate on one single direction through all the data points). The largest outlier appears on FmH2O10 when the B3LYP-D3(0) functional is employed, mostly due to the less accurate MAB-DFPC result for F^- (the error is 0.3 kcal/mol vs. exact SCF) with that functional.

The accurate description of these non-covalent interactions by B97M-V can also be reproduced using MAB-SCF (PC), and the resulting RMS errors (vs. reference data) are compared with the exact SCF results in Figure 2.7. Only minimal differences exist between two sets of RMSDs for the dimer binding energies, while the monotonic deviations of the MAB-SCF (PC) results for each single data point contribute to more pronounced differences for the clusters. Nevertheless, even for H2O6Bind8 where the largest deviation occurs, the RMSD vs. exact SCF results (0.085 kcal/mol) is only 0.1–0.2% of the magnitude of the corresponding binding energies. In terms of the evaluation of relative energies, fairly insignificant errors are caused by using MAB-SCF (PC).

Table 2.11: Binding energies (in kcal/mol) of the L7 complexes evaluated by conventional SCF and MAB-SCF (PC), using two pure density functionals with dispersion corrections (B97-D, B97M-V) and aug-cc-pVTZ as the secondary basis. Unsigned percentage error of the MAB-SCF (PC) results are provided.

	B97-D			B97M-V		
Complex	E_{bind}(exact)	E_{bind}(MAB)	Error (%)	E_{bind}(exact)	E_{bind}(MAB)	Error (%)
C3A	-18.514	-18.505	0.05%	-17.466	-17.488	0.13%
C3GC	-31.120	-31.113	0.02%	-31.050	-31.130	0.26%
C2C2PD	-22.355	-22.343	0.05%	-22.316	-22.329	0.06%
GCGC	-15.302	-15.295	0.05%	-15.460	-15.514	0.35%
GGG	-2.485	-2.530	1.81%	-2.330	-2.392	2.66%
CBH	-15.335	-15.330	0.03%	-12.396	-12.394	0.02%
PBH	-23.311	-23.290	0.09%	-25.867	-25.872	0.02%

Finally we test the accuracy of MAB-SCF (PC) on large non-covalent complexes, using the seven dispersion-bound systems of the recently proposed L7 dataset [277], including (the abbreviations simply follow Ref. 277): stacked circumcoronene-adenine dimer (C3A), stacked circumcoronene with a Watson-Crick G-C base pair (C3GC), parallel displaced coronene dimer (C2C2PD), stacked Watson-Crick G-C base pairs (GCGC), stacked guanine trimer (GGG), parallel stacked octadecane dimer (CBH), and phenylalanine residue trimer (PHE). Due to the tremendous computational effort required for these systems, our calculations were performed with two pure functionals (B97-D, B97M-V) using aug-cc-pVTZ as the secondary basis. Binding energies of these complexes evaluated by conventional SCF and MAB-SCF (PC) are compared in Table 2.11. For most of them, the differences between the results given by two SCF schemes are much smaller than 1% of the magnitude of their binding energies, which indicates the satisfactory accuracy of MAB-SCF (PC) for these large-scale non-covalent interacting systems. The only exception is the guanine trimer (GGG), which is primarily due to its very weak binding energy (-2.33 kcal/mol by B97M-V/aTZ). Due to the lack of reliable reference values for the time being, we cannot meaningfully assess the intrinsic functional errors for L7.

2.6 Discussion and Future Work

The tests performed on a broad range of systems demonstrate that the overall accuracy of this new approximate SCF method seems to be encouraging. As the CBS limit is approached, the functional RMSDs for several representative TC and NC datasets (G2, W4-11, S22) evaluated by MAB-SCF (PC) show very small differences compared to the exact SCF results, as demonstrated in Figures 2.5, 2.6, and 2.7. In particular, "B97M-V/QZVPPD/MAB+PT2" (or "B97M-V/aQZ/MAB+PT2") turns out to be a promising model chemistry, because of

B97M-V's very good accuracy for both TC and NC, its moderate computational cost as a semi-local functional, and its compatibility with the MAB+PT2 approach.

The MAB-SCF (PC) procedure is similar to that of dual-basis methods (including DB-SCF and DFPC), since they both first compute an SCF solution in a primary basis (preferably a subset of the secondary basis) then correct it using perturbation theory. It is encouraging that MAB-SCF (PC) is able to achieve comparable accuracy to DB-SCF with a primary basis adaptively prepared on the fly, whose size is also much smaller than those well-trained DB-SCF basis subsets [215] when the secondary basis approaches the CBS limit. Some comparisons of these two approaches and related discussion are provided in Appendix A.2.4.

At least three challenges remain in successfully applying the MAB-SCF (PC) method. First is the challenge of hybrid density functionals. According to Table 2.2, there is no degradation in terms of the quality of the MAB when a hybrid functional is used: B3LYP in fact has the smallest RMSD for MAB-SCF among the three tested functionals. The reason for the poor performance of MAB+PT2 for hybrid functionals should therefore reside in the PT2 correction represented by Eq. 2.29. Assuming that the MABs calculated by two functionals are of similar quality, the numerators in Eq. 2.29 should also have fairly similar sizes. Thus, functional-dependent differences in the magnitude of the PT2 correction will be largely determined by the denominators (orbital energy differences) in Eq. 2.29. It turns out that the smaller gaps calculated by pure functionals help reduce the errors of MAB-SCF more effectively, while hybrid functionals that usually give larger orbital energy gaps undercorrect. This numerical result is intriguing since the gaps computed by the hybrids are usually deemed to be physically more correct. Although the accuracy of MAB+DFPC is satisfactory in most tested cases, an extra Fock build using the density matrix in the secondary basis (*not* MAB-representable) is required, which would be preferable to avoid (the reason will be elucidated below). Additionally, based on Table 2.6, the accuracy of MAB+DFPC slightly degrades for RSH functionals as well despite its clear advantage over MAB+PT2.

A second limitation is that hypervalent molecules are treated separately in this work by modifying the MAB dimensions in a semi-automated way (choosing to use the "adding vector" approach is user-specified). This is because our algorithm and its related thresholds were chosen to attain the desired accuracy for hypervalent molecules. This option increases the size of the MAB (and thus the computational cost of MAB-SCF) and potentially degrades the convergence of MAB optimization (see the discussion in Appendix A.2.3). Further work on wisely adjusting the algorithm parameters, or possibly refining the present scheme is desirable.

The third general challenge is efficient computational implementation. Although this paper is mainly about the formulation MAB-SCF (PC) and the validation of its accuracy, the ultimate goal is to serve as an inexpensive substitute for conventional SCF as the CBS limit is approached. Compared to normal SCF, our method involves far fewer degrees of freedom. The cost of a single density matrix update, if achieved by diagonalizing the Fock matrix, can be reduced by up to a factor of $(n/m)^3$, where n is the size of the secondary basis, and m is the size of the PRB or MAB. For example, given a model system $(CH_2)_n$, there

Table 2.12: Comparison of the sizes (N) and overlap matrix condition numbers (λ) of the MAB and the target aug-cc-pVTZ basis set on the L7 complexes.

	N(MAB)	N(aTZ)	λ(MAB)	λ(aTZ)
C3A	343	3473	12.17	5.07×10^{12}
C3GC	393	4002	12.57	7.93×10^{12}
C2C2PD	264	2760	13.50	4.32×10^{11}
GCGC	210	2208	9.02	1.39×10^{8}
GGG	180	1863	8.55	1.25×10^{8}
CBH	256	3404	15.63	1.45×10^{9}
PHE	267	3036	13.87	6.24×10^{8}

will be 129 basis functions per –CH_2– unit if def2-QZVPPD is employed as the secondary basis, which is roughly 6 times as large as the PRB (22 functions per unit), and about 18 times as the MAB (7 functions/unit). Therefore, the prefactor of the computational cost of this cubic scaling step is significantly reduced in our scheme. A linear equation solve in the secondary basis dimension is still required to obtain the PT2 correction, while that is far less expensive than diagonalization.

Alternatively, $\mathcal{O}(N)$ scaling electronic structure methods can be potentially applied to the MAB-SCF step, since the overlap matrix of the MAB is extremely well-conditioned, as demonstrated by Table 2.12. The diagonalization-free density matrix update algorithms introduced by Ref. 200 could be used for the MAB-SCF. However, the feasibility of these methods for PRB-SCF is not so clear because one set of diffuse functions is contained in 6-31+G(d).

In practice, however, even for systems as large as the L7 complexes, the Fock matrix construction step still dominates the computational cost of each SCF cycle due to its large prefactor when high accuracy is sought. This is despite the fact that asymptotically that step scales quadratically [278] (or even linearly with special algorithms) [188–195] with respect to system size. In our pilot implementation, the Fock matrix is built from PRB- or MAB-representable density matrices that are of secondary basis dimensions, which costs almost the same as a Fock build in a conventional SCF calculation within the secondary basis. Therefore, to speed up these Fock build steps by taking advantage of the properties of the PRB and MAB becomes the most urgent task for MAB-SCF (PC) to outperform conventional SCF in terms of computational efficiency, especially for medium-sized systems. PRB-SCF can be reformulated as a conventional SCF calculation within a basis set whose size and shell structure are identical to 6-31+G(d), due to the elimination of high angular momentum functions during the basis set projection procedure, and the cost of the involved Fock builds thereby can be significantly reduced without requiring more sophisticated techniques. For MAB-SCF, instead of using Eq. 2.30, the MAB Fock matrix can be constructed by forming the ERI tensor in the primary basis first then contracting it with the MAB density matrix,

Table 2.13: Preliminary timing results (in seconds) for MAB+PT2 on a single processor. The calculations are performed with HF, and aTZ is employed as the secondary basis. The implementation for PRB-SCF is unchanged except that a truncated secondary basis (functions higher than the d shell are excluded) is used. The MAB-SCF is performed with an in-core algorithm, where the MAB ERIs denoted as $(\alpha\beta|\gamma\delta)$ is constructed only once (with RI) and then kept in memory. Two Fock builds are *not* yet optimized ($\mathbf{F}_{\text{ref}}$ for the optimization of MAB and the Fock matrix built upon converged MAB density for PT2) and are still computed based on Eq. 2.30, which dominate the computational costs. With this not yet fully optimized implementation, the MAB-based scheme outperforms conventional SCF by a factor of 4 (on average).

	ethene dimer $(C_2H_4)_2$	benzene C_6H_6	naphthlene $C_{10}H_8$	pyrene $C_{16}H_{10}$
PRB-SCF	35.39	98.31	809.01	4334.50
Build $\mathbf{F}_{\text{ref}}$	86.34	185.09	1397.1	5909.14
MAB opt	3.28	3.34	25.40	63.83
Build MAB ERIs	4.66	7.46	32.99	120.82
SCF in MAB	0.04	0.11	0.80	5.11
Fock build for PT2	85.51	186.33	1400.11	5913.51
PT2	0.20	0.27	0.96	3.05
Total (MAB)	215.42	480.91	3666.37	16349.96
Total (std SCF)	1031.36	1849.79	15323.08	64855.50
Speedup	4.79×	3.85×	4.18×	3.97×

as

$$\mathcal{F}_{\alpha\beta} = h_{\alpha\beta} + 2\,(\alpha\beta|\gamma\delta)\,\mathcal{P}^{\gamma\delta} - \kappa\,(\alpha\gamma|\beta\delta)\,\mathcal{P}^{\gamma\delta} + (V_{xc})_{\alpha\beta}. \tag{2.31}$$

Furthermore, quantities like $(\alpha\beta|\gamma\delta)$ can be efficiently computed (and stored) using the RI approximation, due to the highly compact and atom-blocked structure of the MAB. Preliminary timing results obtained most recently with a not fully optimized implementation (Table 2.13), to a large extent, verify the possibility to speed up SCF calculations with the MAB scheme. We intend to present an optimized implementation of MAB-SCF (PC) and the resulting timings vs. conventional SCF as soon as the above-mentioned challenges are adequately addressed.

2.7 Conclusion

In this work, we proposed a new minimal adaptive basis (MAB) SCF method that can be used in KS-DFT calculations. Its objective is to permit approach to the complete basis set (CBS) limit without explicitly performing the calculation in very large AO basis sets. The key aspects of this paper can be summarized as follows:

- An MAB is obtained as a molecule-adapted, atom-blocked transformation from the secondary AO basis. We have developed a viable optimization method to obtain an MAB using an inexpensive energy-like surrogate function based on a reference SCF calculation in a projected basis of moderate size (the PRB). Compared to exact energy optimization with respect to the transformation (the PAO method), our MAB yields similar total energies with far fewer convergence issues.

- A preconditioned L-BFGS algorithm that requires the gradient and the on-diagonal blocks of the Hessian of the objective function is implemented to solve the MAB optimization problem. In addition, an approach that modifies the MAB dimension based on chemical environment is proposed, and demonstrated to be necessary for hypervalent molecules. These ideas can potentially be used in PAO calculations and other adaptive basis models as well.

- Perturbation corrections (PC) are applied to MAB-SCF to approach the desired accuracy. This resembles DB-SCF without the need to select or develop the paired basis subset. The resulting accuracy is assessed on numerous TC and NC datasets. Measured against exact SCF results, MAB-SCF (PC) generates < 0.15 kcal/mol RMSDs for most of the tested TC datasets, and even smaller errors (usually < 0.1 kcal/mol) for the NCs. Encouragingly, as the CBS limit is approached, the MAB-SCF (PC) method deviates from full SCF one or two orders of magnitude less than the inherent errors in today's best functionals.

- Future work includes further refining the MAB-SCF (PC) model, and developing an efficient implementation, as discussed in Sec. 2.6. We note that hybrid functionals require a different PC to achieve adequate accuracy, and that the strategy to treat hypervalent species is not yet fully automated. Nonetheless, based on the accuracy demonstrated here, and the potential computational advantages of using an atom-centered minimal basis, we believe that this approach merits further development. We hope to report further progress in due course.

Chapter 3

EDA in an "Adiabatic" Picture

3.1 Introduction

Intermolecular interactions play an important role in modulating the structural, vibrational, and other properties of molecular systems. A textbook example is the change in the hybridization of the boron atom from sp^2 to sp^3 upon formation of the NH_3–BH_3 complex. Another is the shift in the IR signal for the stretch of CO when it coordinates with transition metals. [279–281] The origins of many other intriguing phenomena are still under debate. For example, while the nature of the hydrogen bonding interaction in the equilibrium water dimer has been uncovered by many recent studies, [97, 113, 120, 282–285] the main reason underlying the near-linear structure of the hydrogen bond (O–H··O) is still a subject of debate. [120, 285] The microscopic origin of shifts in molecular properties associated with intermolecular interactions is of interest in areas such as the structure of biomolecules, the design of functional materials, the development of empirical force field models, etc.

Energy decomposition analysis (EDA) [92, 97, 104, 105, 108–113, 115, 116, 286–292] methods (which are reviewed in Ref. 98) are useful tools for unraveling the origin of intermolecular interactions. Despite the non-uniqueness of the definition of energy components such as permanent electrostatics (ELEC), Pauli repulsion (PAULI), polarization (POL), charge transfer (CT) and dispersion (DISP), these terms are physically meaningful and represent the correct physics in the asymptotic region as long as the EDA scheme is well designed. Some terms are also naturally combined, such as ELEC+PAULI+DISP which defines the frozen interaction (FRZ), and POL+CT which defines all orbital relaxation (ORB). A representative scheme, which will also be the focus of this paper, is the absolutely localized molecular orbital (ALMO)-EDA [97, 115, 116], yielding

$$\Delta E_{\text{bind}} = \Delta E_{\text{gd}} + \Delta E_{\text{frz}} + \Delta E_{\text{pol}} + \Delta E_{\text{ct}}, \tag{3.1}$$

where ΔE_{gd} is the energy associated with the distortion of monomer (fragment) geometries (GD) upon complex formation. In general, other successful EDA schemes also provide definitions for these components. More importantly, these terms are of high descriptive power

and can be conveniently applied to provide insights into both standard and non-standard intermolecular interactions. As a result, modern EDAs have become very widely applied.

Nonetheless, today's most popular EDA schemes certainly have some remaining gaps in capabilities. For instance, EDAs are typically performed at complex geometries that are obtained from unconstrained electronic structure theory calculations. Two main examples are the equilibrium (optimized) geometry of the intermolecular complex and snapshots from a molecular dynamics simulation. For strong intermolecular interactions, close intermolecular contacts driven by POL and particularly CT often lead to large, repulsive FRZ interactions, which offers little insight apart from obviously indicating considerable interfragment overlap. It would be convenient if such large repulsions could be re-partitioned. It is also true that conventional EDA calculations consist of several energy evaluations at a given geometry with the goal of partitioning a "single-point" interaction energy. Therefore the influence of these physically meaningful components (especially POL and CT) on molecular structures, multipole moments, vibrational frequencies, etc. is *not* directly characterized. A differently formulated EDA method that overcomes these shortcomings is desirable.

One important feature of non-perturbative EDA approaches originating from the Kitaura-Morokuma (KM)-EDA [104, 105] is the use of intermediate states (wavefunctions) to compute energies that are strict upper bounds to the true energy. Energy differences between these states yield components such as FRZ, POL, and CT in Eq. (3.1). Specifically the ALMO-EDA and the block-localized wavefunction (BLW)-EDA [110–112] both use the (non-optimized) antisymmetric product of isolated-fragment wavefunctions as the *initial* or *frozen* wavefunction to compute a frozen energy. For the polarized state, they then variationally optimize an AO-to-MO coefficient matrix that is constrained to be fragment-blocked. Other approaches, such as density constraints,[293, 294] have also been proposed [113, 114]. Well-defined intermediate states allow evaluation of observables including potential energy surfaces (PESs) with different physical content. For instance, POL and CT are both forbidden in the frozen wavefunction, only CT is forbidden in the polarized wavefunction, and there are no constraints in the fully relaxed wavefunction.

The idea of evaluating observables using intermediate EDA wavefunctions at a single geometry dates back to early work by Yamabe and Morokuma [295] and by Stevens and Fink [287], where the influence of different energy components (exchange, POL, and CT) on the dipole moment of the water dimer was investigated. Geometry optimizations on the "CT-forbidden" PES were explored more recently. Weinhold and Klein [296] reported the geometry of the $NH_3 \cdots NH_4^+$ complex optimized on a "resonance-free" PES constructed by the natural bond orbital (NBO) [297] definition. Using the BLW-EDA, Mo and co-workers studied bond length changes due to conjugation [298] and the origin of directionality of some noncovalent bonds [299] by comparing BLW-optimized geometries and fully relaxed ones. CT-induced changes to vibrational properties have also been explored using a similar approach. [112, 300]

In this work we systematize and generalize the idea of using geometries that are optimal for each intermediate state to define an adiabatic version of the ALMO-EDA. The term "adiabatic" is borrowed from spectroscopy, and likewise suggests that conventional EDA at

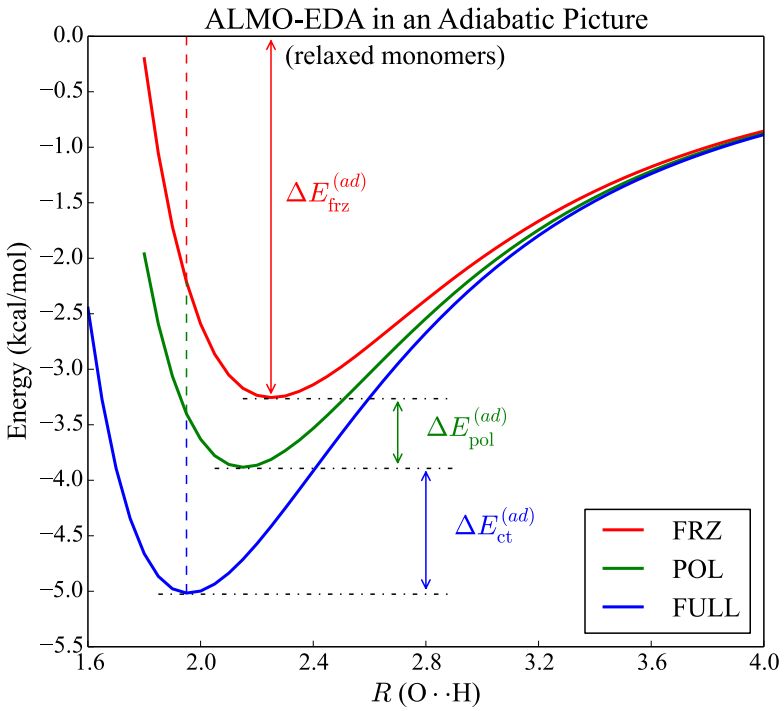

Figure 3.1: Graphical illustration of the adiabatic ALMO-EDA for the case of the water dimer. The three PESs are obtained by relaxing the water dimer geometry while constraining the intermolecular O··H distance (in Å) to be a given value. The zero energy corresponds to the sum of fully relaxed isolated monomer energies. In order to show the difference between the adiabatic and vertical EDAs, the resulting energy components of the adiabatic ALMO-EDA are indicated by double-headed arrows, while the results of vertical ALMO-EDA computed at the fully relaxed geometry are illustrated by the dashed line. Notice that the adiabatic ("*ad*") frozen energy becomes more attractive than in the vertical picture, while CT necessarily becomes less favorable.

a single geometry, as in Eq. (3.1), should be categorized as a "vertical" EDA. The definition of adiabatic contributions is illustrated graphically in Figure 3.1. The procedure begins by accessing the optimal geometry on the frozen PES, and then it is further relaxed by enabling POL and CT sequentially to obtain the final optimized structure. As in the original scheme, the contributions of FRZ, POL, and CT to the total binding energy are additive, so the partitioning of $\Delta E_{\rm bind}$ is simply:

$$\Delta E_{\rm bind} = \Delta E_{\rm frz}^{(ad)} + \Delta E_{\rm pol}^{(ad)} + \Delta E_{\rm ct}^{(ad)}. \tag{3.2}$$

The superscripts distinguish adiabatic contributions from vertical ones, and all three contributions are negative or zero provided that the complex is bound. Note that one can also proceed in the opposite direction, by assessing the fully relaxed geometry first. This will sometimes give a different set of results when multiple stationary points exist on the various

PESs. We will present such examples later in Sec. 3.3. Also, since the geometry of the complex (including the degrees of freedom on individual fragments) is being modified at each stage of this method, there is no GD term in Eq. (3.2). Instead, the contribution of GD is scattered into all three terms in Eq. (3.2). The amount associated with each individual relaxation can be easily identified if desired.

The remainder of this paper is organized as follows. In Sec. 3.2, we present the technical details of the intermediate ALMO-EDA wavefunctions, and for the first time present the analytical gradient of the energy associated with the frozen wavefunction. Combined with the nuclear gradient of the variationally optimized energy under the ALMO constraint (also known as the "self-consistent field for molecular interactions" (SCF-MI) [233, 301] energy), this enables implementation of the adiabatic version of the ALMO-EDA, via Eq. (3.2). We then turn to exploring the insights that can be obtained from this approach by investigating the origin of structural and vibrational frequency changes upon complex formation for a variety of increasingly complicated intermolecular interactions, ranging from water-water and ion-water interactions to metallocenes and the ammonia-borane complex.

3.2 Theory

In the following discussion, we use subscripts to denote the PES on which a specific quantity (energy, density matrix, MOs, etc.) is computed, and superscripts to indicate the PES on which geometry optimization is performed to obtain the associated nuclear configuration. We consider isolated fragments and the frozen, polarized, and fully relaxed PESs for the intermolecular complex, which are denoted by superscripts "0", "frz", "pol", and "full", respectively.

3.2.1 Reformulation of ALMO-EDA in an adiabatic picture

We start by recapitulating the original ALMO-EDA scheme proposed by Khaliullin et al. [115] and summarized in Eq. (3.1). ΔE_{gd} is the energy associated with the distortion of monomer (fragment) geometries (GD) upon the formation of the intermolecular complex:

$$\Delta E_{\mathrm{gd}} = \sum_A E_A^{(\mathrm{X})} - E_A^{(0)}. \tag{3.3}$$

The remaining three terms of Eq. (3.1) are evaluated at a single complex geometry X (which is usually the fully relaxed one) so we will omit the superscripts in Eqs. (3.4)–(3.8), and two intermediate states are involved.

The first intermediate state (the initial or frozen wavefunction) is the antisymmetric product of unrelaxed fragment wavefunctions, which can also be represented by a one-particle density matrix (1PDM) $\mathbf{P}_{\mathrm{frz}}$:

$$\mathbf{P}_{\mathrm{frz}} = (\mathbf{C}_o)_{\mathrm{frz}} \boldsymbol{\sigma}_{\mathrm{frz}}^{-1} (\mathbf{C}_o)_{\mathrm{frz}}^T. \tag{3.4}$$

$(\mathbf{C}_o)_{\text{frz}}$ is the occupied part of the frozen AO-to-MO matrix that is constructed from unrelaxed fragment orbitals, and $\boldsymbol{\sigma}_{\text{frz}}$ is the overlap matrix (metric) of these occupied frozen orbitals:

$$\boldsymbol{\sigma}_{\text{frz}} = (\mathbf{C}_o)_{\text{frz}}^T \mathbf{S} (\mathbf{C}_o)_{\text{frz}}. \tag{3.5}$$

ΔE_{frz} in Eq. (3.1) is then defined as the difference between the energy of this intermediate state and the sum of (deformed) isolated fragment energies:

$$\Delta E_{\text{frz}} = E[\mathbf{P}_{\text{frz}}] - \sum_A E_A. \tag{3.6}$$

The second intermediate state (the polarized wavefunction) is obtained by variationally minimizing the supersystem energy while constraining the AO-to-MO coefficient matrix ($\mathbf{C}$) to be fragment-block-diagonal ("absolutely localized"). As such, only intramolecular relaxation (polarization) is allowed, and charge transfer/delocalization between fragments, at least in principle, does not occur. The contribution of polarization to the interaction is then:

$$\Delta E_{\text{pol}} = E[\mathbf{P}_{\text{pol}}] - E[\mathbf{P}_{\text{frz}}], \tag{3.7}$$

where $\mathbf{P}_{\text{pol}}$ is constructed from $(\mathbf{C}_o)_{\text{pol}}$ (polarized occupied ALMOs) in a similar way as $\mathbf{P}_{\text{frz}}$. Finally, the CT contribution is simply defined as the energy lowering once the ALMO constraint is removed:

$$\Delta E_{\text{ct}} = E[\mathbf{P}_{\text{full}}] - E[\mathbf{P}_{\text{pol}}], \tag{3.8}$$

where $\mathbf{P}_{\text{full}}$ is the 1PDM obtained from a fully relaxed SCF calculation.

Since the two aforementioned intermediate states involve different constraints on their $\mathbf{C}$ matrices, they, together with the fully relaxed state whose energy is optimized without any constraint, belong to different manifolds and represent distinguishable PESs as a function of nuclear configuration. It is possible to locate the minimum on each of these PESs and then evaluate the contribution of FRZ, POL, and CT to the total binding energy in an adiabatic fashion, in contrast to the usual vertical EDA approach. If we initiate the procedure by optimizing the geometry on the frozen PES, the difference between the resulting energy minimum and the sum of undistorted, non-interacting fragment energies is defined as the adiabatic frozen interaction energy:

$$\Delta E_{\text{frz}}^{(ad)} = E[\mathbf{P}_{\text{frz}}^{(\text{frz})}] - \sum_A E_A^{(0)}. \tag{3.9}$$

Both *intramolecular* and *intermolecular* (nuclear) degrees of freedom are relaxed during the geometry optimization above, and the contents of diagonal blocks of the $\mathbf{C}_{\text{frz}}$ matrix also need to be recomputed once the intramolecular nuclear configuration changes. Nevertheless, we still call it the "frozen" PES in the sense that the diagonal blocks of $\mathbf{C}_{\text{frz}}$ are determined by fragment calculations performed in isolation and *not* further relaxed at the supersystem level.

To proceed further, at the optimal geometry on the frozen PES we can "turn on" polarization, which corresponds to a vertical jump onto the polarized PES. We then relax the geometry on this PES until it reaches a local minimum. The stabilization gained in this process relative to the frozen energy at the minimum of its PES defines the contribution of POL in this adiabatic EDA:

$$\Delta E_{\text{pol}}^{(ad)} = E[\mathbf{P}_{\text{pol}}^{(\text{pol})}] - E[\mathbf{P}_{\text{frz}}^{(\text{frz})}]. \tag{3.10}$$

Similarly, the ALMO constraint can then be removed at the optimal geometry on the polarized PES. With another vertical transition followed by another geometry relaxation, the intermolecular complex is brought into its fully relaxed geometry. The adiabatic energy lowering in this last step is then defined as the contribution of CT:

$$\Delta E_{\text{ct}}^{(ad)} = E[\mathbf{P}_{\text{full}}^{(\text{full})}] - E[\mathbf{P}_{\text{pol}}^{(\text{pol})}]. \tag{3.11}$$

This scheme contrasts with the conventional ALMO-EDA where one computes the contribution of each energy component vertically at just the fully optimized geometry.

3.2.2 Analytical gradient on the frozen PES

Analytical nuclear gradients will greatly increase the efficiency of the geometry optimizations needed to locate the minima on the frozen and polarized PESs. The Kohn-Sham (KS)-DFT [5] energy of the complex on the frozen PES can be expressed as (note that E_{frz} and ΔE_{frz} are two distinct quantities)

$$\begin{aligned} E_{\text{frz}} &= E[\mathbf{P}_{\text{frz}}] \\ &= V_{\text{nn}} + \mathbf{P}_{\text{frz}} \cdot \mathbf{h} + \frac{1}{2}\, \mathbf{P}_{\text{frz}} \cdot \text{II} \cdot \mathbf{P}_{\text{frz}} + E_{\text{xc}}, \end{aligned} \tag{3.12}$$

where V_{nn} is the nuclear-nuclear repulsion potential, $\mathbf{h}$ is the core-Hamiltonian, E_{xc} is the exchange-correlation (XC) energy, and II represents the two-electron AO integrals that are used for constructing the Coulomb ($\mathbf{J}$) and exchange ($\mathbf{K}$) matrices. In the most general case,

$$\begin{aligned} \text{II}_{\mu\nu\lambda\sigma} \cdot (\mathbf{P}_{\text{frz}})^{\lambda\sigma} &= [(\mu\nu|\lambda\sigma) + \kappa(\mu\lambda|\nu\sigma)] \cdot (\mathbf{P}_{\text{frz}})^{\lambda\sigma} \\ &= \mathbf{J}_{\mu\nu} + \kappa \mathbf{K}_{\mu\nu}, \end{aligned} \tag{3.13}$$

where the exchange part is scaled by κ ($0 \leq \kappa \leq 1$) based on the proportion of exact exchange contained in the employed functional. The gradient of the energy expressed in Eq. (3.12) with respect to nuclear positions includes the following terms:

$$E_{\text{frz}}^{x} = V_{\text{nn}}^{x} + \mathbf{P}_{\text{frz}} \cdot \mathbf{h}^{x} + \frac{1}{2}\, \mathbf{P}_{\text{frz}} \cdot \text{II}^{x} \cdot \mathbf{P}_{\text{frz}} + E_{\text{xc}}^{x} + E_{\text{frz}}^{\mathbf{S}} \cdot \mathbf{S}^{x} + \sum_{A} (E_{\text{frz}}^{\Delta_A} \cdot \Delta_A^{x}). \tag{3.14}$$

The first four terms are identical to the same components in a standard KS-DFT nuclear gradient except that a special 1PDM is employed. The fifth term, which is related to the

response of the overlap matrix to the displacement of nuclei, has an extra fragment-blocked part:

$$\begin{aligned} E^{\mathbf{S}}_{\mathrm{frz}} \cdot \mathbf{S}^x =& -(\mathbf{PFP}) \cdot \mathbf{S}^x \\ & - \frac{1}{2} \sum_A \left\{ (\mathbf{S}_A^{-1})[(\mathbf{I}-\mathbf{SP})\mathbf{FP}]_A \right\} \cdot \mathbf{S}^x_A \\ & - \frac{1}{2} \sum_A \left\{ [\mathbf{PF}(\mathbf{I}-\mathbf{PS})]_A (\mathbf{S}_A^{-1}) \right\} \cdot \mathbf{S}^x_A \\ =& -\mathbf{W} \cdot \mathbf{S}^x - \sum_A \mathbf{W}'_A \cdot \mathbf{S}^x_A. \end{aligned} \tag{3.15}$$

Note that the subscript "A" indicates that the diagonal block corresponding to fragment A is taken from a global (full-system size) matrix quantity. $\mathbf{F}$ is the KS Fock matrix built from $\mathbf{P}_{\mathrm{frz}}$ (for brevity, the subscripts "frz" have been omitted on the right-hand side (RHS) of Eq. (3.15)). Here the first term is exactly the same as that in a standard SCF gradient, where $\mathbf{W} = \mathbf{PFP}$ is customarily called the energy-weighted density matrix. Similarly, we can define

$$\mathbf{W}'_A = \frac{1}{2} \sum_A \{ (\mathbf{S}_A^{-1})[(\mathbf{I}-\mathbf{SP})\mathbf{FP}]_A + [\mathbf{PF}(\mathbf{I}-\mathbf{PS})]_A (\mathbf{S}_A^{-1}) \}, \tag{3.16}$$

and its expression implies that it is related to the on-fragment coupling between occupied and virtual orbitals through the Fock matrix, which is related to the unrelaxed (frozen) character of $\mathbf{C}_{\mathrm{frz}}$. More details about the derivation of $E^{\mathbf{S}}_{\mathrm{frz}} \cdot \mathbf{S}^x$ are provided in Appendix B.1.1

The last term in Eq. (3.14) is due to the response of fragment MOs to the change in nuclear coordinates, and it is present because the initial wavefunction is not variationally determined. $E^{\Delta_A}_{\mathrm{frz}}$ is the derivative of the supersystem energy with respect to the on-fragment orbital rotations: [117]

$$E^{\Delta_A}_{\mathrm{frz}} = \boldsymbol{\sigma}_A \left[\boldsymbol{\sigma}^{-1} \mathbf{C}_o^T \mathbf{F} (\mathbf{I}-\mathbf{PS}) \mathbf{C}_v \right]_A . \tag{3.17}$$

Note that the matrix quantities on the RHS of Eq. (3.17) are all evaluated on the frozen PES (the subscripts "frz" have also been omitted).

As for many other electronic structure methods, the orbital response contribution to the gradient can be efficiently computed using the z-vector technique: [302]

$$E^{\Delta_A}_{\mathrm{frz}} \cdot \Delta^x_A = \mathbf{z}_A (E^{\Delta_A \mathbf{S}_A}_A \cdot \mathbf{S}^x_A + E^{\Delta_A \mathbf{h}_A}_A \cdot \mathbf{h}^x_A + E^{\Delta_A \mathrm{II}_A}_A \cdot \mathrm{II}^x_A + E^{\Delta_A x}_{\mathrm{xc},A}), \tag{3.18}$$

where $\mathbf{z}_A$ is the solution to the following linear equation:

$$E^{\Delta_A \Delta_A}_A \cdot \mathbf{z}_A = -E^{\Delta_A}_{\mathrm{frz}}. \tag{3.19}$$

The z-vector equation is solved on each individual fragment. $E^{\Delta_A \Delta_A}_A$ is the Hessian matrix of fragment A's KS energy with respect to the occupied-virtual rotation of its MOs, and its RHS is exactly given by Eq. (3.17). More details about solving the z-vector equation and evaluation of the terms contracted with $\mathbf{z}_A$ on the RHS of Eq. (3.18) are provided in Appendix B.1.2.

3.2.3 Analytical gradient on the polarized PES

The energy of the intermolecular complex evaluated on the polarized PES shares the same expression as E_{frz} except that a different density matrix ($\mathbf{P}_{\text{pol}}$) is used:

$$E_{\text{pol}} = V_{\text{nn}} + \mathbf{P}_{\text{pol}} \cdot \mathbf{h} + \frac{1}{2}\, \mathbf{P}_{\text{pol}} \cdot \text{II} \cdot \mathbf{P}_{\text{pol}} + E_{\text{xc}}. \tag{3.20}$$

Nonetheless, its analytical nuclear gradient is much simpler to compute compared to that of E_{frz}. This is because the supersystem energy is variationally optimized at this stage through the SCF-MI procedure, and at convergence it satisfies $E_{\text{pol}}^{\Delta_A} = \mathbf{0}$. Therefore, the contribution from orbital response, as in the last term in Eq. (3.14), vanishes in the nuclear gradient of E_{pol}. Moreover, $[(\mathbf{I} - \mathbf{SP})\mathbf{FP}]_A$ can be rewritten as

$$[(\mathbf{I} - \mathbf{SP})\mathbf{FP}]_A = [(\mathbf{I} - \mathbf{SP})\mathbf{F}\mathbf{C}_o\boldsymbol{\sigma}^{-1}\mathbf{C}_o^T]_A = [(\mathbf{I} - \mathbf{SP})\mathbf{F}\mathbf{C}_o\boldsymbol{\sigma}^{-1}]_A(\mathbf{C}_o^T)_A, \tag{3.21}$$

where the second equality holds due to the fragment-blocked structure of $\mathbf{C}_o$. Since $[(\mathbf{I} - \mathbf{SP})\mathbf{F}\mathbf{C}_o\boldsymbol{\sigma}^{-1}]_A$ vanishes when SCF-MI converges, [117] we also have $\mathbf{W}'_A = 0$ on the polarized PES. Taken together, the nuclear gradient of E_{pol} turns out to be identical to the standard KS-DFT gradient:

$$E_{\text{pol}}^x = V_{\text{nn}}^x + \mathbf{P}_{\text{pol}} \cdot \mathbf{h}^x + \frac{1}{2}\, \mathbf{P}_{\text{pol}} \cdot \text{II}^x \cdot \mathbf{P}_{\text{pol}} - \mathbf{W} \cdot \mathbf{S}^x + E_{\text{xc}}^x. \tag{3.22}$$

We note that the nuclear gradient of the converged SCF-MI energy has been presented by Famulari et al. [303] and Mo [298] before, and it has the same form as Eq. (3.22).

3.3 Results

3.3.1 Computational details

The adiabatic ALMO-EDA presented above is implemented in a locally modified version of the Q-Chem 4.4 package.[255] The original ALMO model [234] is employed to construct the polarized PESs in this work, and to be consistent, the conventional (vertical) ALMO-EDA results are also computed based on the original AO-blocked scheme, [115] but with a further decomposition of the FRZ term. [118] The ωB97X-V functional [72], which employs the VV10 [71] non-local correlation functional (NLC) for the description of dispersion, is used for all calculations. ωB97X-V shows high accuracy for non-covalent interactions so it is an ideal functional to use together with our DFT-based EDA schemes. Results for the H_2O dimer and the $H_2O\cdots Cl^-$ complex computed with other functionals are also provided in Appendix B.2 (Tables B.1–B.3). Unless otherwise specified, the def2-TZVPPD [183] basis (augmented triple-ζ) is employed for all the energy and property (geometry, vibrational frequency) calculations. For numerical integrations involved in DFT calculations, a (75, 302) grid (75 radial shells with 302 Lebedev points in each) is used for the XC functional, and the SG-1 grid [258] is used for the NLC functional.

All obtained stationary points on the various PESs have been characterized using either full-spectrum frequency calculations or the Hessian-free approach introduced by Sharada et al. [304] Since the second functional derivative of VV10 is not available, all vibrational frequencies are computed based on analytical nuclear gradients using a finite-difference approach (the step size is set to 10^{-3} Å). Additionally, by generalizing the approach of Ref. 304, we can compute the contribution of VV10 to the LHS of Eq. (3.19) and the RHS of Eq. (3.18) using finite-difference matrix-vector products, which enable geometry optimizations on the frozen PES with the ωB97X-V functional. Further details are presented in Appendix B.1.3.

3.3.2 Water dimer

We first look at the widely studied water dimer. Its fully relaxed geometry (the global minimum linear configuration) is shown in the top panel of Figure 3.2, where the proton donor of the hydrogen bond is denoted as H_d. The adiabatic ALMO-EDA results are collected in columns 2–4 of Table 3.1, where the adiabatic energy lowerings correspond to the three terms in Eq. (3.2), while E_{bind} refers to the depth of each PES. The largest adiabatic contribution to binding comes from the frozen interaction (65%). A comparison with the vertical ALMO-EDA results at the fully relaxed geometry was already shown in Figure 3.1, indicating that the importance of FRZ increases in the adiabatic picture, as guaranteed by the formulation of this method. By contrast, the magnitude of ΔE_{pol} is reduced by almost one half and that of ΔE_{ct} becomes about 1/3 smaller.

Nonetheless, CT significantly affects the structure and vibrational frequencies of the water dimer, according to the adiabatic EDA results. Compared to the equilibrium geometry on the polarized but CT-forbidden surface, the hydrogen bond is shortened by about 0.2 Å, and the tilting angle of the proton acceptor H_2O (the β angle) is reduced by 6°. The preference for $\beta \approx 127°$ was explained using the complementary occupied-virtual orbital pairs [305] (COVPs) in Ref. 282. CT also results in the slight elongation of the O_1–H_d bond and the remarkable red shift in the symmetric O–H stretch frequency for the H-donor water molecule (ω_9, relative to 3863 cm^{-1} for the H_2O monomer as computed at the same level of theory). In contrast, the frequency of the symmetric O–H stretch for the H-acceptor water molecule (ω_{10}) is minimally affected, which results in a 100 cm^{-1} splitting between them. The effects of CT on the asymmetric O–H stretch frequency of the H-donor water molecule (ω_{11}), on the other hand, are considerably smaller. These findings are fairly consistent with the previous study by Ramos-Cordoba et al. [300] The effect of POL is smaller than that of CT. Besides shortening the $O_2 \cdots H_d$ distance by 0.1 Å, POL has very little effect on the β angle and the vibrational frequencies, which is also consistent with its fairly small energetic contribution.

There has been much discussion of the origin of the near-linear structure of the water dimer. A recent study [285] using symmetry-adapted perturbation theory (SAPT) [92] revealed that the energy difference between the linear configuration and the bifurcated transition structure [306] (middle panel of Figure 3.2) is *not* due to the first-order terms of SAPT (electrostatics and so-called exchange-repulsion). It is the second-order terms (especially induction) that render the linear configuration more favorable. This agrees with the

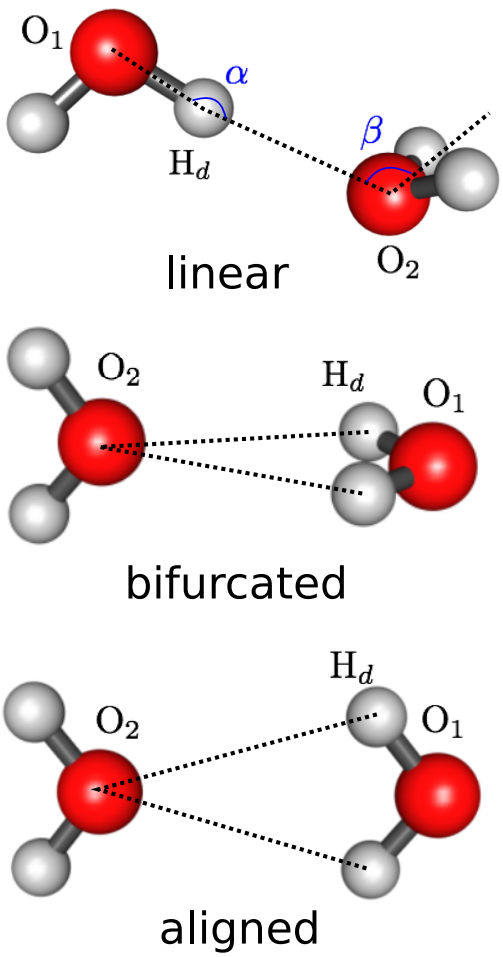

Figure 3.2: Water dimer in the most stable linear configuration (C_s) (top), and two less stable geometries, labeled as "bifurcated" (middle) and "aligned" (lower) (C_{2v}). O_1 is for the oxygen atom in H-donor and O_2 is for that in H-acceptor. For the linear configuration, α denotes $\angle O_1$–H_d–O_2 and β denotes the angle between $O_2 \cdot\cdot H_d$ and the bisector of the H-acceptor water (the tilting angle of the H-acceptor water).

Table 3.1: Adiabatic EDA results for the water dimer in the linear, bifurcated and aligned configurations, as shown in Figure 3.2. The binding energies (in kJ/mol) are calculated with regard to the fully relaxed water monomers, and the adiabatic ΔE's are the three energy components (FRZ, POL and CT). The distances are reported in Å, angles in degrees (°), and vibrational frequencies in cm^{-1}. ω_1 and ω_2 are the two lowest vibrational modes. ω_9 and ω_{10} are the symmetric O–H stretches of the hydrogen donor and acceptor water molecules, respectively, and ω_{11} and ω_{12} correspond to the asymmetric ones.

	linear			bifurcated			aligned		
	FRZ	POL	FULL	FRZ	POL	FULL	FRZ	POL	FULL
Adiabatic ΔE	-13.62	-2.63	-4.74	-11.11	-1.42	-1.06	-8.20	-0.99	-0.58
ΔE_{bind}	-13.62	-16.25	-20.99	-11.11	-12.53	-13.59	-8.20	-9.19	-9.77
$R(O_2 \cdot\cdot H_d)$	2.25	2.15	1.96	2.66	2.60	2.53	2.80	2.73	2.68
$R(O_1 \cdot\cdot O_2)$	3.21	3.11	2.92	3.15	3.08	3.02	3.29	3.22	3.17
$R(O_1$–$H_d)$	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96
α angle	172.66	172.83	173.20	–	–	–	–	–	–
β angle	132.36	132.15	126.22	–	–	–	–	–	–
ω_1	88.31	105.93	125.50	$146.99i$	$155.23i$	$219.31i$	$176.20i$	$217.65i$	$248.22i$
ω_2	97.77	117.15	141.78	50.70	61.49	58.98	$157.31i$	$168.62i$	$179.56i$
ω_9	3861.01	3852.88	3754.43	3861.79	3861.33	3859.58	3862.75	3862.40	3861.65
ω_{10}	3869.45	3861.08	3856.55	3871.77	3872.11	3868.25	3878.45	3875.06	3873.87
ω_{11}	3960.70	3955.07	3937.68	3957.10	3953.65	3945.65	3961.22	3959.00	3955.68
ω_{12}	3965.15	3959.02	3955.20	3962.04	3961.33	3960.32	3963.37	3960.66	3960.07
$\omega_{10} - \omega_9$	8.44	8.20	102.12	9.98	10.78	8.67	15.70	12.66	12.22
$\omega_{12} - \omega_{11}$	4.45	3.95	17.52	4.96	7.68	14.67	2.15	1.66	4.39

ALMO-EDA results obtained with second-order Møller-Plesset perturbation theory (MP2), [120] which demonstrated that POL and CT (corresponding to SAPT's induction term) stabilize the linear geometry, because the aligned configuration (the lower panel of Figure 3.2) is preferred by the frozen interaction (either with or without dispersion). CT, on the other hand, is of predominant importance in NBO analysis, [296, 297] since the water dimer is unbound at its equilibrium geometry if NBO-derived CT is removed. [307] However, many other studies [282, 299, 300, 308] suggested that CT is *not* decisive for the directionality of the water dimer H-bond.

Both the studies of Refs. 120 and 285 rely on EDA calculations performed along a coordinate scan on the fully relaxed PES. As mentioned above, this may result in too unfavorable dispersion-free frozen interaction (the first-order terms in SAPT) and lead to an (over)emphasized role of POL and CT (induction). Here we provide a different perspective by means of the adiabatic ALMO-EDA. The first interesting result from Table 3.1 is that the linear configuration is calculated to be a minimum on the frozen PES, based on its positive lowest frequency (ω_1). Despite the large $O_2 \cdot\cdot H_d$ distance, it is remarkable that $\angle O_1$–H_d–O_2 (the α angle) at this geometry shows little difference compared to that in the fully relaxed geometry. Thus a near-linear structure of the water dimer is favored even without the presence

of POL and CT!

Table 3.2: Vertical ALMO-EDA results (in kJ/mol) for three configurations of the water dimer computed at their fully relaxed geometries (data in the parentheses are computed at the optimal geometries on the frozen PES). DFFRZ (dispersion-free frozen interaction) refers to the sum of ELEC and PAULI, which corresponds to the sum of first-order terms in SAPT.

	linear	bifurcated	aligned
ELEC	-63.68 (-33.93)	-34.92 (-26.72)	-22.51 (-17.85)
PAULI	61.90 (24.75)	29.45 (19.91)	18.51 (12.97)
DFFRZ	-1.78 (-9.08)	-5.47 (-6.81)	-4.00 (-4.87)
DISP	-7.62 (-4.55)	-5.67 (-4.54)	-4.34 (-3.54)
POL	-4.95	-1.74	-1.17
CT	-6.80	-1.16	-0.62
INT	-21.15	-14.05	-10.13

To further verify this result, we explored two other configurations (bifurcated and aligned) of the water dimer as well. The vertical ALMO-EDA results for all three configurations are shown in Table 3.2. On the fully relaxed PES, the sum of ELEC and PAULI (the dispersion-free frozen term) for the bifurcated and aligned structures are indeed more favorable than that of the linear water dimer, and it seems tempting to draw a conclusion that the linear configuration becomes the global minimum eventually thanks to its much stronger stabilization by POL and CT (the contributions from POL and CT are much smaller at bifurcated and aligned structures according to the vertical EDA results). However, a different picture emerges from the adiabatic EDA results in Table 3.1. The linear configuration is energetically favored across the board, even before induction effects (POL and CT) come into play. Also, according to the frozen energy components evaluated on the *frozen* PES (see Table 3.2), the sum of ELEC and PAULI does favor the linear structure instead.

In terms of the adiabatic binding energy at the FRZ level, the linear configuration is *already* 2.5 kJ/mol more favorable than the bifurcated, and 5.4 kJ/mol more favorable than the aligned. Such an energetic ordering is also in accord with the evaluated frequencies of the lowest two vibrational modes (ω_1 and ω_2): the linear configuration is characterized to be an energy minimum on all PESs, while the bifurcated and aligned structures appear to be a transition state (TS) and a second-order saddle point throughout, respectively. The adiabatic ALMO-EDA results therefore show that the near-linear equilibrium geometry of the water dimer is *not* determined by the induction effects (POL and CT) — it is energetically preferred at the FRZ level already. This preference, according to Table 3.2, primarily arises from the more favorable permanent electrostatics at the linear structure, which perhaps goes against expectations since alignment of the molecular dipoles does *not* determine the conformational preference for the water dimer at the FRZ level.

Table 3.3: Adiabatic ALMO-EDA results for the $H_2O\cdots Cl^-$ complex. Results in columns 2–4 (energy minima) are obtained by starting from the optimal geometry on the fully relaxed PES, while those in columns 5 and 6 (TS) are obtained by starting from the optimal geometry on the frozen PES. Under C_s structures, ω_5 corresponds to the stretch of O–H_d and ω_6 corresponds to the stretch of the free O–H bond, while under C_{2v} structures they correspond to the symmetric and asymmetric O–H stretches of the water molecule, respectively. Other details about data presentation are the same as in Table 3.1.

	FRZ (C_{2v})	POL (C_s)	FULL (C_s)	POL (C_{2v})	FULL (C_{2v})
Adiabatic ΔE	-41.31	-9.03	-12.41	-8.68	-6.26
ΔE_{bind}	-41.31	-50.33	-62.74	-49.98	-56.24
R(Cl··O)	3.37	3.32	3.12	3.28	3.14
R(Cl··H_d)	2.84	2.48	2.15	2.75	2.60
R(O–H_d)	0.96	0.97	0.99	0.96	0.97
∠Cl–H_d–O	115.32	145.91	166.71	115.41	115.60
∠H–O–H	99.44	100.11	101.38	98.38	96.57
ω_1	138.89	118.14	192.77	$35.69i$	$312.28i$
ω_5	3869.56	3801.45	3406.22	3853.15	3805.38
ω_6	3911.89	3919.23	3922.35	3891.68	3817.02
split ($\omega_6 - \omega_5$)	42.33	117.78	516.13	38.53	11.64

3.3.3 Water-ion interactions

Water-chloride

The $H_2O\cdots Cl^-$ complex is an archetype for water-anion interactions. Like the water dimer, its equilibrium geometry (the bottom panel of Figure 3.3) has C_s symmetry and a near-linear hydrogen bond between Cl^- and one of the protons (the bond angle in the equilibrium geometry is 166.7°). By excluding CT and POL successively and relaxing the geometry of this system on the polarized and frozen PESs, we obtain two stationary geometries, and the adiabatic ALMO-EDA results along this path are collected in columns 2–4 in Table 3.3. Characterized by the lowest vibrational frequency (ω_1), the two geometries are both energy minima. They are shown in the top two panels of Figure 3.3. Without CT, the Cl··H_d distance is elongated by 0.33 Å, and ∠Cl··H_d–O is decreased by over 20°, which are in reasonable agreement with the results in Ref. 300. The substantial change in ∠Cl··H_d–O due to the removal of CT contrasts with the water dimer case where CT only slightly modifies the hydrogen bond angle (∠O··H_d–O). With the further relaxation on the frozen PES, Cl^- migrates onto the bisector of ∠H–O–H, resulting in a complex with C_{2v} symmetry. This geometry resembles the bifurcated configuration of the water dimer, which, however, is a saddle point on water dimer's frozen PES.

We suppose that the differences between water-water and water-chloride interactions are primarily due to the different characteristics of their permanent electrostatic interactions. On the frozen PES, the linear water dimer is favored by permanent electrostatics (see Table 3.2),

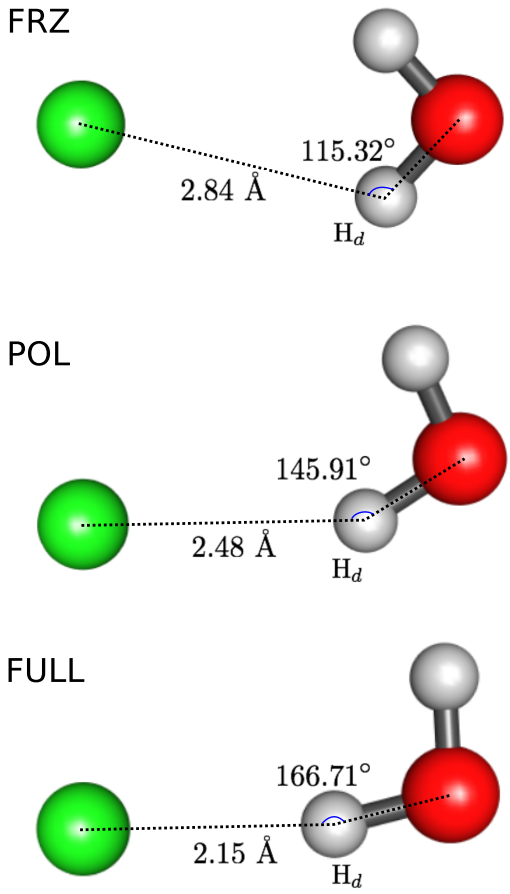

Figure 3.3: Optimized geometries (energy minima) of the $H_2O\cdots Cl^-$ complex on the frozen, polarized, and fully relaxed PESs. Changes in $R(Cl{\cdot\cdot}H_d)$ and $\angle Cl–H_d–O$ are indicated.

as opposed to the "classical" intuition that dipole-dipole interaction should prefer the bifurcated structure. The frozen interaction energy thus serves as the initial driving force for the formation of its near-linear structure. In contrast, for $H_2O\cdots Cl^-$ the monopole-dipole-like electrostatic interaction favors the bifurcated (C_{2v}) configuration, and CT, which has larger magnitude in this system, compensates for the diminished stabilization from ELEC and results in the near-linear $\angle Cl{\cdot\cdot}H_d–O$ in the fully relaxed structure. It should be noted that dispersion, despite its relatively small magnitude, also helps stabilize the C_{2v} structure, since use of the B3LYP functional [34, 35, 42] (which is almost dispersion-free) results in a broken-symmetry structure for the energy minimum on the frozen PES as a trade-off for less unfavorable geometry distortion, while the addition of a D3 correction [57] to B3LYP restores the C_{2v} symmetry of that minimum-energy structure (see Tables B.2 and B.3 in Appendix B.2 for more details).

Compared to the O–H stretch frequencies of a fully relaxed water monomer (3863 and

3965 cm^{-1}), the frequency of the asymmetric mode in the frozen complex (ω_6) is red-shifted by over 50 cm^{-1}. This might be related to the deformation of the water molecule in this complex, whose bond angle is substantially reduced from 105.06° (for a fully relaxed monomer) to 99.44° (such a change increases the dipole moment of H_2O and thus enhances the attractive electrostatic interaction). To verify that, we compute the vibrational frequencies of an isolated water molecule with the same geometry as in the frozen complex, and the two resulting O–H stretch frequencies (3856 and 3914 cm^{-1}) are in good agreement with ω_5 and ω_6 in the complex.

Once the C_{2v} symmetry is broken, ω_5 and ω_6 correspond to the stretch of the O–H_d and another free O–H bond, respectively. POL results in a red-shift in the O–H_d stretch frequency by 68 cm^{-1}, while for the free O–H bond, the vibrational frequency is blue-shifted and becomes slightly closer to the higher-energy mode in a free water molecule, which further enlarges the splitting between ω_5 and ω_6. The O–H_d bond is more significantly weakened when CT is enabled, which is usually described as an $n \rightarrow \sigma^*$ donation. Its impact is manifested in the 0.02 Å elongation of the O–H_d bond and the substantial red shift (400 cm^{-1}) in ω_5 relative to their values evaluated on the previous PES (polarized but CT-forbidden). Since the stretch of the free O–H bond is almost unaffected, we see a 4-fold increase for the splitting between ω_5 and ω_6.

Table 3.4: Vertical ALMO-EDA results (in kJ/mol) for $H_2O\cdots Cl^-$ computed at the obtained stationary geometries on the frozen and polarized PESs. The contribution from CT is excluded from the total interaction energy; and for the stationary point on the frozen PES, POL is also excluded.

	Frozen (C_{2v})	Polarized (C_s)	Polarized (C_{2v})
ELEC	-78.71	-87.96	-89.45
PAULI	43.92	55.37	55.50
DISP	-8.54	-9.22	-9.76
FRZ	-43.33	-41.82	-43.70
POL	–	-10.17	-9.15
INT	-43.33	-51.99	-52.85

By starting from the optimal geometry on the frozen PES and relaxing it on the polarized and fully relaxed PESs sequentially, the C_{2v} symmetry of the complex will be conserved if no other perturbation to the geometry is exerted, and thereby we end up with a new adiabatic path where Cl^- lies on the bisector of ∠H–O–H throughout. Indicated by their imaginary ω_1's in Table 3.3, the resulting structures on the polarized and fully relaxed PESs are both TSs that connect the minima on two sides of the bisector. Nevertheless, it is notable that on the polarized PES, the two stationary geometries (C_s and C_{2v}) have extremely close adiabatic binding energies — the C_s configuration is marginally more favorable by 0.35 kJ/mol. We investigate these two structures (as well as the optimal geometry on the frozen PES) using

the vertical ALMO-EDA, where the energy components that are not allowed on the PES each structure is associated with are excluded, and the results are shown in Table 3.4. It is striking that in terms of the interaction energy (evaluated without relaxing the monomers), the TS (C_{2v}) structure is *more favorable* than the stable C_s structure, although the latter has a more favorable polarization energy; and the C_s structure becomes the minimum simply because the water molecule has a less distorted geometry. Note that the order of these two energetically "near-degenerate" stationary points on the polarized PES can be altered by using a different, less accurate density functional (see Table B.3 in Appendix B.2). This is no different from the need to wisely choose the model chemistry in any standard chemical study using DFT.

Similar to the scenario of the bifurcated (or aligned) water dimer, both the energy magnitude of CT and its effect on observables are diminished when Cl^- is located on the bisector, since the associated donor-acceptor orbitals are of less favorable orientations. [300] The symmetric and asymmetric O–H stretches are both red-shifted by CT (and marginally by POL), and the splitting between them decreases upon the transition from frozen to fully relaxed PES. This is related to the reduction in ∠H–O–H when Cl^- approaches the H_2O molecule, which, very interestingly, causes ∠Cl··H_d–O to be almost constant. Here we see a nuanced balance between favorable intermolecular interactions (ELEC, and probably POL and CT as well) and the penalty from distorted monomer geometries.

Water-divalent cations

We next consider the complexes formed by H_2O and divalent alkali earth metal cations (Mg^{2+} and Ca^{2+}), whose geometries are shown in Figure 3.4. These interactions are regarded as challenging cases for the development of polarizable force fields, [167, 309, 310] and for this purpose it will be interesting to characterize the impact of CT on energetics and molecular properties of these systems.

The adiabatic ALMO-EDA results for water-Mg^{2+} and water-Ca^{2+} complexes are collected in Table 3.5. All optimized geometries have C_{2v} symmetry. In terms of energetics, FRZ has the largest contribution to the total binding energy owing to the strongly attractive permanent electrostatics. The role of POL is much more pronounced in these systems

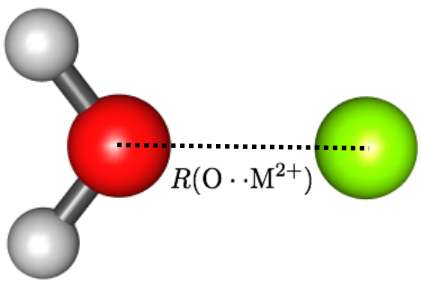

Figure 3.4: Geometry of the water-divalent cation (Mg^{2+}, Ca^{2+}) complexes (C_{2v})

Table 3.5: Adiabatic ALMO-EDA results for the $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ complexes. The absolute (abs) shifts of ω_5 (symmetric O–H stretch) and ω_6 (asymmetric O–H stretch) are evaluated relative to the vibrational frequencies of these two modes in a fully relaxed water monomer ("+" for blue shifts and "-" for red shifts), and the relative (rel) shifts are evaluated with respect to the vibrational frequencies computed at the minimum of the previous PES. Other details about data presentation are the same as in Table 3.1.

	$H_2O\cdots Mg^{2+}$			$H_2O\cdots Ca^{2+}$		
	FRZ (C_{2v})	POL (C_{2v})	FULL (C_{2v})	FRZ (C_{2v})	POL (C_{2v})	FULL (C_{2v})
Adiabatic ΔE	-176.69	-151.37	-12.70	-129.44	-80.37	-26.99
ΔE_{bind}	-176.69	-328.06	-340.77	-129.44	-209.81	-236.80
R(O··M^{2+})	2.08	1.93	1.92	2.50	2.35	2.24
R(O–H)	0.96	0.98	0.98	0.96	0.97	0.97
∠H-O-H	106.44	105.33	105.65	104.67	103.85	104.31
ω_5	3896.01	3672.81	3663.60	3870.37	3746.09	3708.35
ω_6	3936.50	3708.79	3700.98	3921.58	3791.05	3754.73
rel shift (ω_5)	–	223.20	9.21	–	124.28	37.74
rel shift (ω_6)	–	227.71	7.81	–	130.53	36.32
abs shift (ω_5)	+32.77	-190.43	-199.64	+7.13	-117.15	-154.89
abs shift (ω_6)	-28.74	-256.45	-264.26	-43.66	-174.19	-210.51

compared to that in the hydrogen-bonding complexes discussed above, due to the presence of divalent cations. For $H_2O\cdots Mg^{2+}$, the adiabatic ΔE_{pol} has nearly comparable magnitude to ΔE_{frz}, while for $H_2O\cdots Ca^{2+}$, $\Delta E_{\text{pol}}/\Delta E_{\text{frz}}$ is close to 2/3 since the former decays faster with the increase of intermolecular distance. The energetic contribution of CT is fairly insignificant (1/30 of the total binding energy) for the $H_2O\cdots Mg^{2+}$ complex, but it is of more importance for $H_2O\cdots Ca^{2+}$ (> 10% of the total binding energy) despite its larger O··M^{2+} separation.

In both systems, the equilibrium O··M^{2+} distances are shortened by 0.15 Å due to the effect of POL. The rearrangement of water's electron density due to the strong polarization effect appears to weaken the two O–H bonds, which are elongated by 0.02 Å for $H_2O\cdots Mg^{2+}$ and 0.01 Å for $H_2O\cdots Ca^{2+}$. Accompanying the increased bond lengths are the red shifts in the harmonic frequencies of symmetric (ω_5) and asymmetric (ω_6) O–H stretches. While CT has a non-trivial effect on the properties of the $H_2O\cdots Ca^{2+}$ complex (shortening the O··M^{2+} distance by an additional 0.11 Å and red-shifting the two O–H stretch modes by over 35 cm^{-1}), it only makes a minimal difference to $H_2O\cdots Mg^{2+}$. This is consistent with the relative energetic importance of CT in these two systems.

Several other subtle features are evident in Table 3.5. For instance, relative to a free water molecule (105.06°), ∠H–O–H is slightly expanded in $H_2O\cdots Mg^{2+}$ but reduced in $H_2O\cdots Ca^{2+}$ (on all three PESs). We also notice that for both systems, the symmetric O–H mode is blue-shifted while the asymmetric one moves to red on the frozen PES (relative to their frequencies in free H_2O), which reduces the gap between their vibrational frequencies. These changes are not solely the effect of geometry distortion: for an isolated water molecule

Table 3.6: Adiabatic ALMO-EDA results for four metallocenes (of D_{5d} symmetry): $FeCp_2$ (singlet), $NiCp_2$ (triplet), $CuCp_2$ (doublet), $ZnCp_2$ (singlet). Details about data presentation are the same as in Table 3.1.

	$FeCp_2$			$NiCp_2$		
	FRZ	POL	FULL	FRZ	POL	FULL
Adiabatic ΔE	-1794.19	-642.36	-597.67	-1825.04	-671.59	-292.66
ΔE_{bind}	-1794.19	-2436.54	-3034.21	-1825.04	-2496.64	-2789.30
R(M–C_5)	2.23	2.04	1.66	2.17	2.00	1.83
$\angle C_5$–H	3.49	0.80	-1.41	3.23	0.53	-0.54
	$CuCp_2$			$ZnCp_2$		
	FRZ	POL	FULL	FRZ	POL	FULL
Adiabatic ΔE	-1839.19	-699.33	-245.36	-1849.27	-706.44	-45.52
ΔE_{bind}	-1839.19	-2538.52	-2783.89	-1849.27	-2555.71	-2601.22
R(M–C_5)	2.15	1.98	1.91	2.13	1.97	1.95
$\angle C_5$–H	3.08	0.35	0.13	3.01	0.33	0.43

of the same geometry as in the frozen $H_2O\cdots Mg^{2+}$ complex, the two O–H stretch frequencies are 3846 and 3959 cm^{-1}, which are fairly different from their values evaluated in the complex. These changes might be related to the distortion of electron densities upon the formation of the intermolecular complex, as an effect of Pauli repulsion, and further investigations of these intriguing subtleties could be worthwhile.

3.3.4 Metallocenes

Unraveling the complicated nature of metal-ligand bonding is of great importance in organometallic chemistry. Ferrocene ($FeCp_2$, Cp = cyclopentadiene) and other metallocenes are systems that exhibit strong dependence of bonding on the identity of the metal cation. Using the generalized Ziegler-Rauk EDA, [106–109] Rayón and Frenking pointed out the relative importance of electrostatics and orbital interactions (POL and CT) in ferrocene, and showed that Cp→Fe π-donation makes the largest contribution to bonding. [311] Swart reported a study of many metallocenes with the same EDA approach, where issues like the most stable spin state and the relative importance of electrostatics vs. orbital interactions for each metallocene were investigated. [312]

In this work, we apply the adiabatic ALMO-EDA to four staggered (D_{5d}) metallocenes formed by first-row transition metals: $FeCp_2$, $NiCp_2$, $CuCp_2$ and $ZnCp_2$. Their geometries and the structural parameters to be investigated are shown in Figure 3.5 (using ferrocene as an example). The most favorable spin states for $FeCp_2$ and $NiCp_2$ were taken from Refs. 312 and313, and were verified with ωB97X-V/def2-TZVPPD as well. In order to see the trends as a function of metal from Fe^{2+}, whose $3d$ orbitals are roughly half-filled, to the fully filled

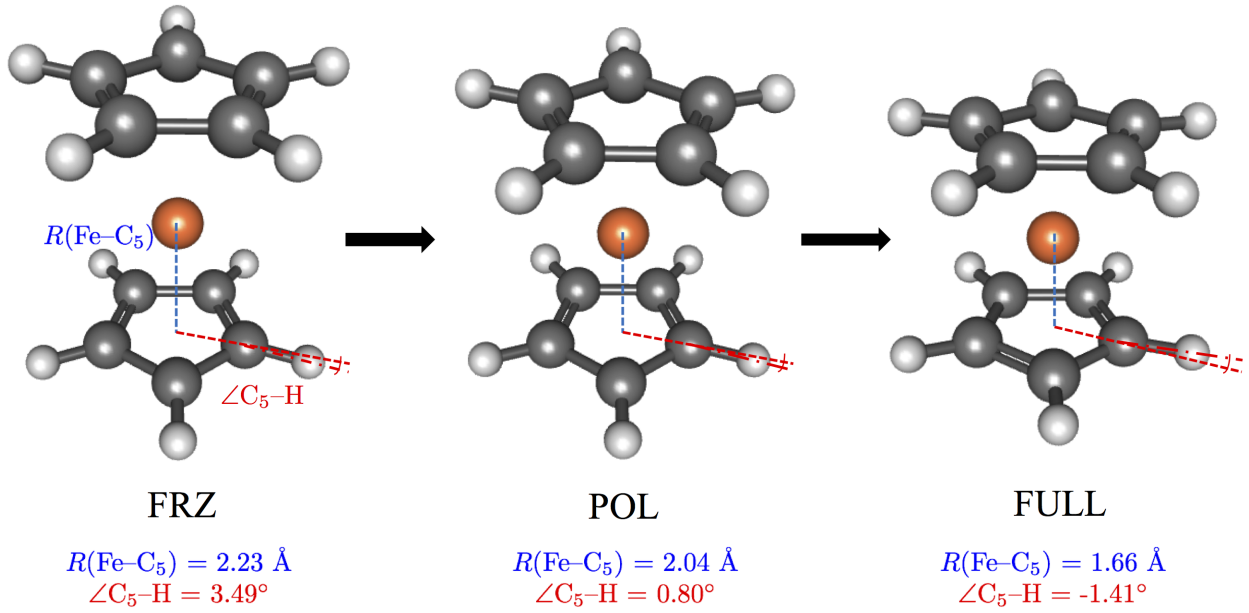

Figure 3.5: Illustration of the changes in optimal structures of a staggered (D_{5d}) ferrocene upon the transition from frozen to fully relaxed PES. The geometric parameters investigated in this work are indicated in the figure. $\angle C_5$–H is the angle between the C_5 plane and one C–H bond (represented by red dash-dot lines), which is defined to be *positive* when the bending direction of C–H is *away from* the metal cation. The gradually shortened Fe–C_5 distance and the decreased and finally flipped $\angle C_5$–H upon the transition are also demonstrated, and their values (shown at the bottom) are identical to those in Table 3.6. For other metallocenes studied in this work, the same figure can also help visualize the changes in their geometries.

Zn^{2+}, we always keep the spin state of the metal cation the same as the most stable one at the fully relaxed geometry, i.e., the energy consumed by the spin-flip of Fe^{2+} (from a high-spin isolated cation to a closed-shell complex) is *not* included.

As shown in Table 3.6, under the adiabatic picture, the largest stabilization is gained at the FRZ stage for all four metallocenes, which is reasonable because of the strongly attractive permanent electrostatic interaction between M^{2+} and two Cp^- ligands. From Fe^{2+} to Zn^{2+}, the magnitude of ΔE_{frz} slightly increases (spanning a fairly small range of 55 kJ/mol) and the equilibrium M–C_5 distance on the frozen PES is gradually reduced. This trend arises from contraction of the ionic radius from Fe^{2+} to Zn^{2+}, rendering the latter systems less prone to Pauli repulsion with closer metal-ligand contacts. This picture is supported by decomposition of the frozen interaction energies at their optimal geometries on the frozen PESs (Table 3.7). Use of the geometries optimized on the frozen PES provides a clearer trend than the frozen energy decomposition at fully relaxed geometries. The contributions from PAULI and DISP vary minimally, despite the decreasing M–C_5 distance from $FeCp_2$ to $ZnCp_2$. Thus the increasingly stabilizing FRZ term mainly reflects the enhancement of permanent electrostatics. One notable feature of these optimal initial structures is the

Table 3.7: Decomposition of the vertical frozen interaction energies (in kJ/mol) of four metallocenes evaluated at the optimized geometries on their frozen PESs.

	$FeCp_2$	$NiCp_2$	$CuCp_2$	$ZnCp_2$
ELEC	-1989.64	-2018.75	-2030.08	-2041.30
PAULI	253.42	254.11	250.56	253.11
DISP	-63.07	-65.26	-64.30	-65.67
FRZ	-1799.29	-1829.90	-1843.82	-1853.86

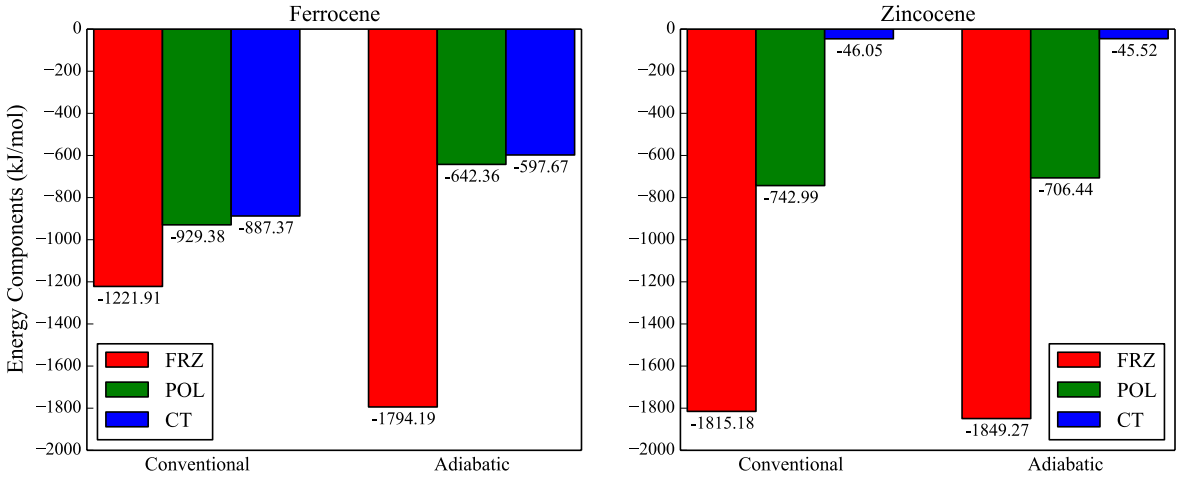

Figure 3.6: Comparison between the conventional (vertical) and adiabatic ALMO-EDA results for $FeCp_2$ (left) and $ZnCp_2$ (right). The results for the conventional ALMO-EDA are computed at the fully relaxed geometries obtained at the ωB97X-V/def2-TZVPPD level of theory, where the GD terms are 4.44 and 2.99 kJ/mol for these two systems, respectively, which are rather small and not plotted in the figure.

outward bending of the C–H bonds, which are favored by permanent electrostatics due to the generation of local dipoles (pointing outwards) on the C_5 rings. The bending angle decreases with the shortening of the M–C_5 distance.

Polarization also makes a significant contribution to the binding of these systems, driven by the proximity of a metal dication and the considerably polarizable, electron-rich aromatic ligands. The resulting adiabatic energy lowerings (ΔE_{pol}) are in a range between -600 and -700 kJ/mol (marginally above 1/3 of the magnitude of ΔE_{frz}), and the M–C_5 distances are shortened by a similar amount (0.15–0.20 Å) for all four investigated complexes. Moreover, in terms of the magnitude of polarization energies, the equilibrium M–C_5 distances, and the C_5–H angles, a similar trend emerges from $FeCp_2$ to $ZnCp_2$ as on the frozen PES.

The energetic contribution of CT decreases from $FeCp_2$ to $ZnCp_2$, indicating that less full $3d$ orbitals facilitate the donation from Cp's π-electrons. For $FeCp_2$, ΔE_{ct} constitutes roughly

20% of the binding energy. Its adiabatic CT energy is roughly 300 kJ/mol smaller than the vertical counterpart evaluated at the fully relaxed geometry (see the left panel of Figure 3.6). Indeed, the discrepancy between the energetics given by vertical and adiabatic EDAs will be most pronounced when CT is most significant, since the adiabatic $\Delta E_{\rm ct}$ incorporates the energy consumed to overcome the penalty of Pauli repulsion when moving towards the equilibrium geometry on the fully relaxed PES.

For ferrocene, there is a substantial shortening of the Fe–C_5 distance after CT is allowed. While it reflects the important role of CT in this system (twice as large as the shortening due to POL), it also enlarges the gap between the vertical and adiabatic CT energies. By contrast, for $NiCp_2$ the magnitude of $\Delta E_{\rm ct}$ is reduced by 50% relative to that for $FeCp_2$, and the shortenings of the Ni–C_5 distance owing to POL and CT are about equal. This shows the diminished effect of CT when the $3d$ orbitals are occupied with two more electrons. Turning to the fully filled Zn^{2+}, the CT effect becomes completely insignificant (the Zn–C_5 distance is shortened by only 0.02 Å) so that the vertical and adiabatic approaches yield very similar results for this system (see the right panel of Figure 3.6).

We note that the Fe–C_5 distance and $\angle C_5$–H of our fully optimized ferrocene (staggered) are in good agreement with a reference geometry optimized at the CCSD(T)/TZV2P+f level of theory. [314] Also, for all four metallocenes, the structural parameters obtained at the full SCF level are fairly close to the previously reported values [312] computed with OPBE/TZP [315], although the metallocenes investigated in Ref. 312 were exclusively in the eclipsed conformation.

One feature of $FeCp_2$ and $NiCp_2$ is that the C_5–H angles are flipped when CT is enabled, i.e., their C–H bonds are bent towards the metals instead (as illustrated in Figure 3.5). While the final value of $\angle C_5$–H is clearly an indicator of the significance of CT, it does not seem fully correct to assert that inversion of the C–H bending directions is a direct consequence of CT. For these systems, the value of $\angle C_5$–H has already substantially decreased when POL is enabled, although none of them has changed sign at that stage. We think that the inversion of C–H bending directions is induced by the closer contact between M^{2+} and Cp^- in the fully optimized geometries of $FeCp_2$ and $NiCp_2$, which is driven by the relatively large effect of CT in these systems.

3.3.5 Ammonia-borane complex and basis set dependence

The NH_3-BH_3 complex exhibits a strong dative (CT) interaction, associated with the coupling of a Lewis base, NH_3, with a Lewis acid, BH_3, which is planar in isolation. The adiabatic ALMO-EDA results are listed in Table 3.8, and show very large changes in R(N··B) and $\angle$N–B–H (the latter characterizes pyramidalization of BH_3) due to different components of this interaction. The structure of NH_3, on the other hand, is much more rigid ($\angle$B–N–H changes by less than 1°). At the optimal geometry on the frozen PES, $\angle$N–B–H deviates only slightly from 90°, as a result of the balance between enhanced permanent electrostatics and the penalty from increased geometry distortion. POL makes a rather small energetic contribution that is similar to the size of $\Delta E_{\rm frz}$. However, POL results in a remarkable

change in the complex geometry, as the N··B distance and ∠N–B–H are roughly mid-way between their values at the frozen and the fully relaxed structures. CT, as we expect, is the dominant component of this interaction. In the adiabatic picture, it stabilizes the complex by 103 kJ/mol, which is over 3 times larger than the sum of ΔE_{frz} and ΔE_{pol}. It further shortens the N··B distance by 0.44 Å, and expands ∠N–B–H to 105° (characteristic of sp^3 hybridization).

Table 3.8: Adiabatic ALMO-EDA results for the NH_3-BH_3 complex (of C_{3v} symmetry). Details about data presentation are the same as in Table 3.1.

	FRZ	POL	FULL
Adiabatic ΔE	-15.64	-14.83	-103.25
ΔE_{bind}	-15.64	-30.47	-133.72
R(N··B)	2.76	2.10	1.66
∠N–B–H	92.08	98.38	104.97
∠B–N–H	111.56	111.26	110.79

The adiabatic EDA provides a significantly different picture than the vertical ALMO-EDA (Figure 3.7). Due to the close contact between NH_3 and BH_3 in the final structure, the vertical ALMO-EDA shows a large geometric distortion (GD), a strongly repulsive FRZ term and a very large value of POL that exceeds the magnitude of CT. We note that the latter might be partly due to the use of the original ALMO model to separate POL and CT (*vide infra*). The adiabatic EDA provides complementary and perhaps deeper insight into this interaction. Although ΔE_{ct} is about 20 kJ/mol smaller than the vertically evaluated value, it is absolutely dominant compared to other energy components, and FRZ and POL are both moderately attractive because of the relatively large intermolecular separations in geometries optimized on the frozen and polarized PESs.

One shortcoming of the original ALMO-EDA (and the related BLW-EDA) is the basis set dependence of the energy associated with the polarized (AO-blocked SCF-MI) state. [103, 117, 316] When very diffuse basis sets are employed, basis functions on one fragment can partially describe CT to another fragment even with the fragment-blocked ALMO coefficient matrix. The resulting polarization energy can therefore be lowered due to the contamination of CT. In this adiabatic version of the ALMO-EDA, the effect of POL on the complex geometry and other properties might also be overemphasized. This overestimation is usually considered to be insignificant when a basis set no larger than augmented triple-ζ is employed.[117]

It is possible that more serious errors arise for the NH_3-BH_3 complex because its interaction energy is so strong and so CT-dependent. As a first test, Figure 3.8 shows the geometries of NH_3-BH_3 optimized on the polarized PES with three triple-ζ basis sets: def2-TZVPP [183], def2-TZVPPD, and aug-cc-pVTZ (aTZ) [178, 179] (with increasing number of basis functions). The resulting geometries are rather sensitive to the choice of basis set:

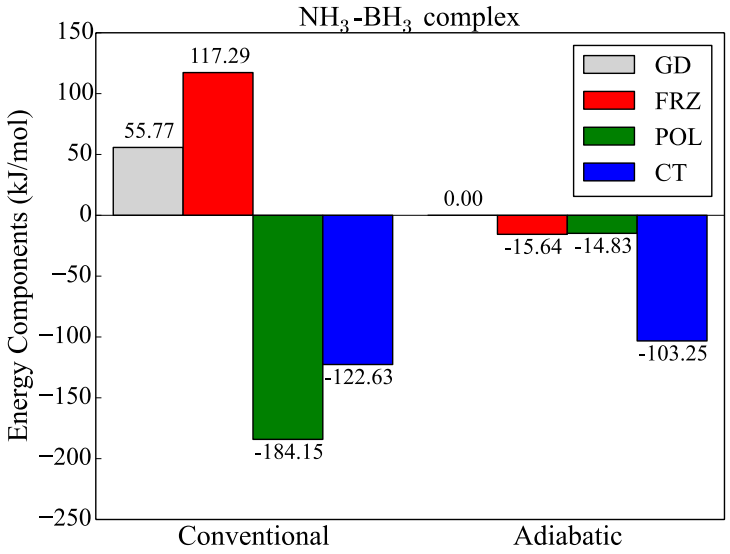

Figure 3.7: Comparison between the conventional and adiabatic ALMO-EDA results for the NH_3–BH_3 complex (of C_{3v} symmetry). The results of the conventional (vertical) ALMO-EDA are computed at the fully relaxed geometry optimized at the ωB97X-V/def2-TZVPPD level of theory.

while the structure optimized with def2-TZVPP is closer to that of the frozen complex, the one optimized with aTZ is more similar to the fully relaxed geometry, as can be seen from the values of R(N··B) and ∠N–B–H. (Note: on the frozen and fully relaxed PESs, basis set dependence of the resulting structures is much weaker, which is confirmed by Table B.4 in Appendix B.2.) Although def2-TZVPPD (the basis set employed in the present work) and aTZ are both of augmented triple-ζ quality, the POL-induced shortening of the N··B distance and pyramidalization of BH_3 are less exaggerated in the resulting geometry of def2-TZVPPD.

The recently proposed fragment electrical response function (FERF) model [117] is an improved method for the evaluation of the polarization energy, which gives it a non-trivial basis set limit. Despite the analytical nuclear gradient of SCF-MI with FERFs being unavailable for the time being, thanks to the small size of this system, it is possible to use finite differences to optimize the NH_3-BH_3 geometry. The resulting SCF-MI (FERF) geometries computed with the above three triple-ζ basis sets are shown in Figure 3.9. Neither R(N··B) nor ∠N–B–H is significantly affected by changing the basis set, demonstrating the much weaker basis set dependence of the FERF model. In addition, the POL-induced changes of the complex geometry are less prominent than those manifested in the results obtained by using the original ALMO model. While non-trivial, formulation and implementation of the analytical gradient of the SCF-MI (FERF) model is therefore highly desirable for a clean separation of the effects of POL and CT in the adiabatic ALMO-EDA.

For completeness, we also examine the basis set sensitivity of several key equilibrium intermolecular distance parameters obtained on the polarized PESs for other systems that

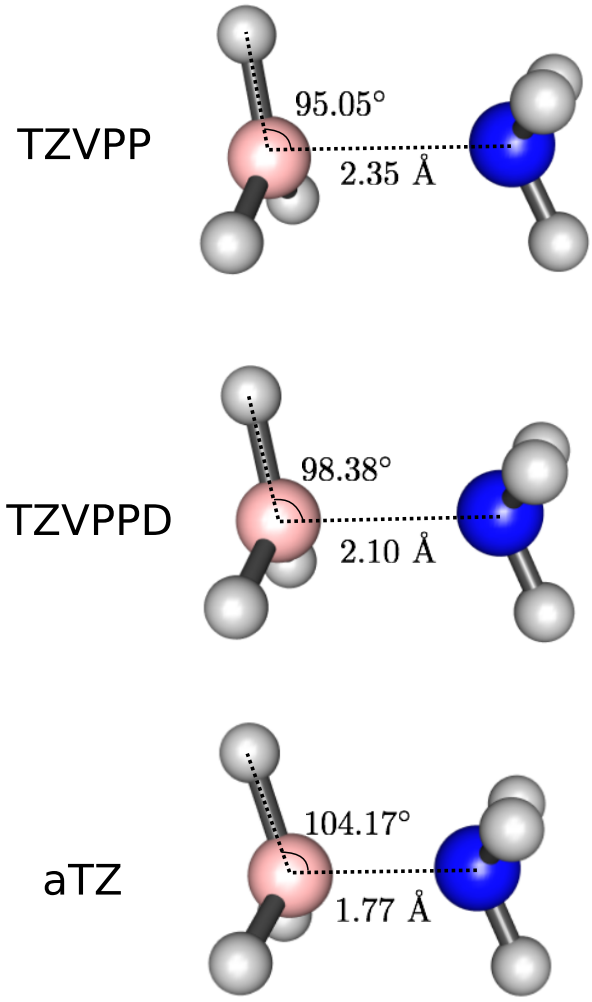

Figure 3.8: Basis set dependence of the equilibrium geometry on the polarized PES constructed with the SCF-MI (ALMO) model. The exhibited geometries are optimized with def2-TZVPP, def2-TZVPPD, and aug-cc-pVTZ from top to bottom.

are investigated in this work. The results generated with the same set of triple-ζ bases are listed in Table 3.9. In general, the basis set dependence of these intermolecular distances is much less pronounced compared to the NH_3–BH_3 case. None of these changes would qualitatively affect the interpretation of the results, and for most of them, the fluctuations are smaller than 0.05 Å. Note that using a larger basis set (in terms of number of basis functions) does not necessarily result in shorter intermolecular distances on the polarized PES, as can be seen in Table 3.9.

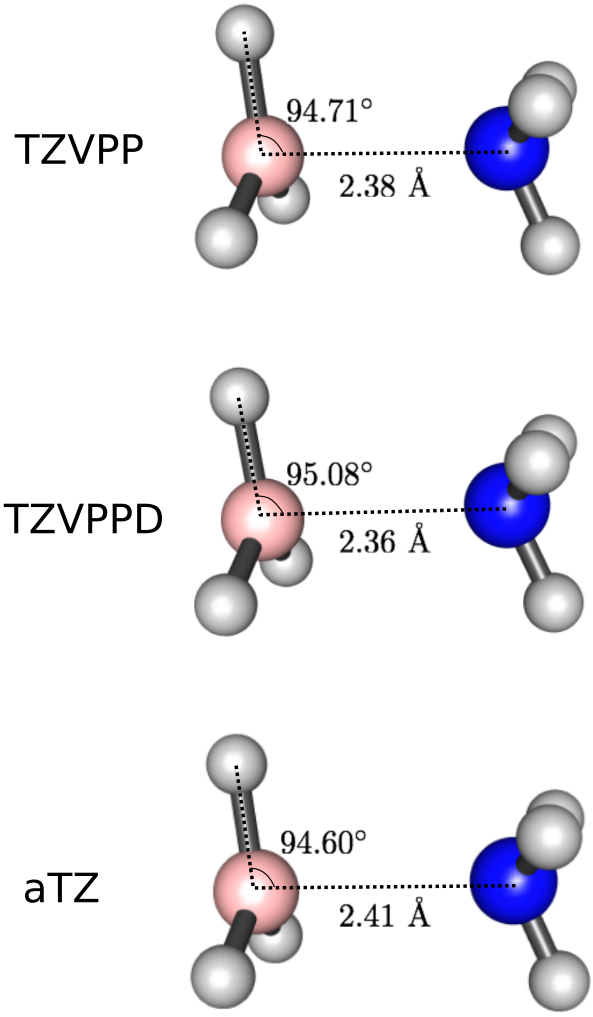

Figure 3.9: Basis set dependence of the equilibrium geometry on the polarized PES constructed with the SCF-MI (FERF) model. The exhibited geometries are optimized with def2-TZVPP, def2-TZVPPD, and aug-cc-pVTZ from top to bottom.

3.4 Conclusion

In this work, by contrast with conventional EDA methods that decompose a "vertical" interaction energy at a given nuclear configuration, we advocate the concept of an "adiabatic EDA" and reformulate the popular ALMO-EDA within that picture. While this method retains the main properties of the original ALMO-EDA (e.g., a variational separation of POL and CT), it has the following new features:

- The energy components are defined by the *adiabatic* energy differences between the PESs that correspond to the three intermediate (including the final) states involved in the

Table 3.9: Basis set dependence of the key intermolecular distance parameters (in Å) for systems other than the NH_3–BH_3 complex that are investigated in this work. The geometries are optimized on the polarized PES using the original ALMO model, and results generated by three triple-ζ basis sets are compared.

System	Distance	TZVPP	TZVPPD	aTZ
water dimer	$R(\mathrm{O}\cdot\cdot\mathrm{H}_d)$	2.12	2.15	2.12
water-Cl^-	$R(\mathrm{Cl}\cdot\cdot\mathrm{H}_d)$	2.39[a]	2.48	2.45
water-Mg^{2+}	$R(\mathrm{O}\cdot\cdot\mathrm{Mg}^{2+})$	2.03	2.04	2.08
water-Ca^{2+}	$R(\mathrm{O}\cdot\cdot\mathrm{Ca}^{2+})$	2.34	2.35	– [b]
$FeCp_2$	$R(\mathrm{Fe}$–$\mathrm{C}_5)$	2.03	2.04	2.08

[a] Basis sets that contain diffuse functions should be necessary for systems that contain anions.
[b] aug-cc-pVTZ basis for Ca is unavailable.

ALMO-EDA, which are the frozen, polarized, and fully relaxed PESs. Accessing the minimum (or other stationary points) on each PES is required to evaluate the energy components, and this is facilitated by analytical nuclear gradients (the nuclear gradient associated with the frozen PES, to the best of our knowledge, is reported here for the first time).

- The resulting energy components differ from those obtained by a conventional (vertical) ALMO-EDA calculation performed at the fully relaxed geometry. By definition, the adiabatic EDA yields a more favorable frozen interaction energy, whereas the energetic contribution of CT is reduced (that of POL is usually reduced as well). This circumvents the often unfavorable FRZ term in the conventional ALMO-EDA so that the resulting frozen energy and its decomposition are easier to interpret (e.g., Table 3.7). On the other hand, for systems where the effect of POL or CT is strong, the adiabatic EDA still manifests their significance. In some cases, such as NH_3-BH_3, the adiabatic picture seems more helpful in providing a physical interpretation (e.g., Figure 3.7).

- By relaxing the geometry on each PES successively, the modulation of structural properties of the intermolecular complex due to POL and CT and even the often nuanced effect of FRZ can be isolated. Furthermore, other molecular properties (in this work we focused on vibrational frequencies) can be computed at the stationary points on each PES, showing how different components of an intermolecular interaction affect these properties. This connects the components of the adiabatic EDA scheme more directly to experimental observables.

Illustrated by proof-of-concept applications to a fairly broad range of systems, we feel that the adiabatic EDA is useful in providing a complementary and perhaps more "panoramic"

view of intermolecular interactions than a conventional EDA at a single geometry. The adiabatic EDA provides more comprehensive information relating to both the origin and the implications for observables of each component of the interaction. The adiabatic EDA approach, in principle, can be applied to many other EDA methods (e.g., the density-based EDA [113] by Wu et al.) as long as the nuclear gradient for each intermediate state is available. Also, beyond structure and vibrational frequencies, the impact of intermolecular interactions on many other electronic structure related molecular properties (e.g., electronic excitation energies, NMR chemical shifts, etc.) can be explored with this method. We hope to report such applications in due course.

Chapter 4

On the Definition of Charge Transfer in EDA

4.1 Introduction

The charge-transfer (CT) phenomenon, in the context of intermolecular interaction, usually refers to the stabilization effect arising from the intermolecular relaxation of the system wavefunction. As indicated by its name, CT is often accompanied with changes in the charge populations of the binding moieties. Given a complex consisting of an electron donor (D) and acceptor (A), CT can also be viewed as a partial mixing of the higher-energy ionic state $|D^+A^-\rangle$ into the neutral state $|DA\rangle$, which thus lowers the ground state energy of the system. [94] Note that CT discussed in this paper is a phenomenon accompanied with formation of ground state complexes, which is essentially a delocalization effect due to the quantum nature of electrons, and it differs from other widely studied concepts such as electron-transfer (ET) reactions [317, 318] where integral number of electron(s) is/are transfered. Therefore, the term "charge delocalization", instead of "charge transfer", was suggested by Misquitta [102] to refer to this phenomenon.

As an important binding force besides electrostatic interaction and dispersion, the effect of CT is ubiquitous in non-covalently bound complexes. For systems where the CT effect is non-trivial, the magnitude of the associated stabilization energy ranges from a few kJ/mol in simple hydrogen-bonding complexes, to hundreds of kJ/mol in strong dative compounds. CT also plays an important role in radical chemistry, since an open-shell species can serve as either electron donor or acceptor. [319–322] Beyond energetic effects, CT reveals itself via abundant observable consequences, such as changes in structural and vibrational properties of the binding moieties. One well-recognized example is the elongation of the X–H bond and the red shift in the X–H stretch frequency upon the formation of the X–H$\cdots$Y hydrogen bond. [296, 323] Nonetheless, the CT energy itself is not an observable in quantum mechanics, so it is almost impossible to measure the stabilization and other effects directly in experiment, although a series of molecular beam scattering experiments that extract the CT component

indirectly based on several assumptions have been reported. [324–326]

Theoretical models of CT, on the other hand, are valuable tools for characterizing this important effect in intermolecular binding. They also facilitate the efforts to incorporate the CT component explicitly in advanced molecular force fields. [144, 327–330] Most of these models are associated with energy decomposition analysis (EDA) [98] schemes, and they often rely on the preparation of a "CT-free" reference state, on which CT between molecules is forbidden subject to a certain criterion. The CT energy is then evaluated with a subtraction ($\Delta E_{\text{CT}} = E_{\text{full}} - E_{\text{CT-free}}$) or using perturbation theory. Therefore, the characteristics of a given CT model are implicitly determined by its definition of this intermediate state.

The natural bonding orbital (NBO) method [297, 331] constructs a set of mutually orthogonal occupied orbitals (the "Lewis" orbitals, including core, lone-pair and bonding orbitals) from a density matrix derived from a fully converged self-consistent field (SCF) calculation, whereby the reference state is defined. The CT stabilization energy is then defined as the interaction between donor's bonding/lone-pair orbitals and acceptor's anti-bonding orbitals via the Fock operator, which can be calculated by deleting the corresponding off-diagonal blocks in the Fock matrix (see Ref. 331 for details). Another closely related approach is the natural energy decomposition analysis (NEDA) scheme, [289, 290, 332] where the CT term is defined as the difference between the energy of the fully relaxed supersystem wavefunction and that of the state constructed from the "Lewis" orbitals. These approaches are often found to generate excessively large CT energies, and a recent investigation by Stone [333] has shown that the "Lewis" orbitals obtained from the NBO procedure, which are *not* variationally optimized, provide a rather poor description for the supersystem energetically, rendering the resulting CT energy significantly contaminated by basis set incompleteness error.

Symmetry-adapted perturbation theory (SAPT), [92, 334] by its original formulation, does not separate out CT as an individual term. The intra- and intermolecular orbital relaxation due to the presence of other molecules, i.e., polarization (POL) and CT, both belong to the induction term in SAPT. Several methods have been proposed to extract CT energy from the induction energy of SAPT. It was first suggested that the CT energy can be evaluated as the difference in second-order induction (and exchange-induction) energy calculated with dimer-centered (supersystem) basis and monomer-centered basis.[101, 335] However, it is evident that CT under this definition vanishes when monomer basis becomes complete. To overcome the drawbacks of this approach, Misquitta proposed an alternative method [102] under the framework of SAPT(DFT), [336] which employs a regularized (damped in the short range) Coulomb potential for fragment B's nuclei to suppress the $A \to B$ charge transfer. The CT energy at the second-order of SAPT is then defined as

$$E^{(2)}_{\text{CT}} = [E^{(2)}_{\text{ind}} - E^{(2)}_{\text{ind}}(\text{Reg})] + [E^{(2)}_{\text{ind-exch}} - E^{(2)}_{\text{ind-exch}}(\text{Reg})] \tag{4.1}$$

Note that the total second-order CT energy is the sum of the $A \to B$ and $B \to A$ contributions. This method will be further discussed in Sec. 4.3.7.

In contrast to the NBO and SAPT-based methods, many other EDA schemes originating from the Kitaura-Morokuma (KM)-EDA [104, 105] access the "CT-free" state via a varia-

tional optimization. It is most straightforward to separate POL and CT through a partition in the orbital space, i.e., given fragments A and B, the mixing of A's virtual orbitals (V_A) into its own occupied orbitals (O_A) in the presence of B is defined as the polarization of A, while the rotation of O_A into $V_A \oplus V_B$ with O_B remaining frozen further incorporates the $A \to B$ CT. This is exactly how early EDA approaches such as the reduced variational space (RVS) [287] and the similar constrained space orbital variation (CSOV) [286] methods are formulated. However, as distinct "CT-free" states are employed to obtain the $A \to B$ and $B \to A$ contributions, the CT (and POL) terms computed thereby do not add up to the full variational energy lowering, and thus higher-order relaxation effect is not incorporated.

The later proposed absolutely localized molecular orbital (ALMO)-EDA [97, 115, 116] (and the closely related block-localized wavefunction (BLW) [110–112] method) overcomes this problem, by employing a single "CT-free" state for the entire system. Starting from the supersystem wavefunction constructed with isolated fragment molecular orbitals (MOs), the polarized yet "CT-free" intermediate state is obtained through a variational optimization subject to the constraint that the MOs on each fragment are only expanded by in its own atomic orbital (AO) basis functions, i.e., the AO-to-MO coefficient matrix has a fragment-block-diagonal structure. The CT energy is then defined as the energy lowering due to the further relaxation of the system wavefunction upon the removal of the ALMO constraint:

$$\Delta E_{\mathrm{CT}} = E[\Psi_{\mathrm{full}}] - E[\Psi_{\mathrm{ALMO}}] \tag{4.2}$$

Note that the basis set superposition error (BSSE) can be removed from the right-hand side (RHS) of Eq. (4.2) if desired, by applying a counterpoise correction [88]. An ALMO-based charge-transfer analysis (CTA) that utilizes the same "CT-free" state has also been proposed, [305] which is able to decompose the transfered electron density (ΔQ), as well as the associated ΔE_{CT}, into pairwise forward and backward contributions. The ALMO-EDA and CTA have been successfully applied to quantifying CT in small water clusters [337], naphthalene-benzene radical cation complexes, [321] ionized glycerol complex [338], etc. More details about this method, including recent development to achieve a meaningful basis set limit for the POL and CT components, are introduced in Sec. 4.2.1.

The density-based EDA (DEDA), [113] on the other hand, relies on a population-based definition of CT: the "CT-free" state is prepared by performing a constrained density functional theory (CDFT) [293, 294, 339] calculation that variationally optimizes the supersystem wavefunction while conserving charge population of each fragment to a preset value. Real-space partitioning methods, such as the Becke [34] and Hirshfeld [340] schemes, are found to be compatible with the CDFT approach. While we refer the reader to Sec. 4.2 for more details, we note that several desirable features of the CDFT definition have been shown/argued recently: (i) the resulting CT energy (ΔE_{CT}) is stable with respect to basis set, and by construction it is BSSE-free; [103, 341] (ii) E_{CT} evaluated with CDFT shows better linear correlation with the amount of transferred charge (Δq) and the reciprocal of the IP(donor)-EA(acceptor) gap compared to other methods [103] (IP: ionization potential; EA: electron affinity). It has also been noticed that CDFT usually yields smaller CT energies

Table 4.1: CT stabilization energy (in kJ/mol) for the equilibrium water dimer reported in literature. Note that the difference in optimized water dimer geometries and employed model chemistries amongst these studies is partially responsible for the discrepancy in CT energies.

Method	CT energy	Reference
NBO	-38.37	343
NEDA	-37.28	290
KM-EDA	-8.83	104
	-5.19	344 [a]
RVS	-3.56	287
ALMO-EDA (I)	-7.3	282
ALMO-EDA (II) [b]	-7.74	97
CDFT(Becke)	-2.8	113
CDFT(FBH) [c]	-3.31	341
IMPT [d]	-3.49	335
SM09 [e]	-2.65	101
Regularized SAPT	-1.49	102

[a] The method in Ref. 344 is essentially the KM scheme with BSSE corrected.
[b] (I)/(II) refers to the first-/second-generation ALMO-EDA.
[c] The definition of "Becke" and "FBH" can be found in Sec. 4.2.2.
[d] IMPT stands for intermolecular perturbation theory [345].
[e] "SM09" stands for the method reported in Ref. 101 under the framework of SAPT(DFT).

than ALMO or BLW approaches, [103] and the role of CT is thus less emphasized in the resulting energy decomposition. [342]

The lack of a unique definition of CT in theoretical models has made it "*the source of a good deal of error and confusion*" (Stone). [94] For the water dimer, a prototypical system for intermolecular hydrogen bonding, the magnitude of CT stabilization energy produced by different methods varies significantly (collected in Table 4.1; also see Table 1 in Ref. 307), leading to inconsistent interpretations of the nature of this hydrogen bond. While the NBO analysis clearly suggests that CT is predominant since the water dimer is unbound without CT at its equilibrium geometry, the ALMO-EDA and other variational methods based on orbital-space partitioning suggest no more significant role of CT than other components such as frozen interaction (FRZ) and POL, and even smaller CT energies are given by the CDFT and regularized SAPT methods.

The goal of the present paper is to at least partly resolve the confusions caused by distinct definitions of CT. With a broad range of simple model systems, we elucidate the

essential differences between definitions based on orbital-space partitioning and those based on population changes, using ALMO and CDFT as respective representatives. Moreover, the recently proposed "adiabatic EDA" approach [119] is utilized to demonstrate the differences in calculated observable consequences of CT subject to these two definitions, which aims to assess their capability of characterizing well-established phenomena of CT in chemistry. Other related aspects, such as the differences between the results of ALMO and regularized SAPT, are also discussed in this paper.

4.2 Methods

4.2.1 ALMO-based definition of CT

The ALMO approach defines the "CT-free" state through a partition in the orbital space. It first constructs the frozen wavefunction of the complex by concatenating the fragment wavefunctions optimized in isolation with full antisymmetrization. The associated one-particle density matrix (1PDM), $\mathbf{P}_{\text{frz}}$, can be represented as

$$\mathbf{P}_{\text{frz}} = (\mathbf{C}_o)_{\text{frz}}(\boldsymbol{\sigma}_{\text{frz}}^{-1})(\mathbf{C}_o)_{\text{frz}}^T, \tag{4.3}$$

where $(\mathbf{C}_o)_{\text{frz}}$ is an AO-to-MO coefficient matrix containing occupied fragment MOs on its diagonal blocks (as illustrated in Fig. 4.1), and $\boldsymbol{\sigma}_{\text{frz}}$ is the overlap between these orbitals:

$$\boldsymbol{\sigma}_{\text{frz}} = (\mathbf{C}_o)_{\text{frz}}^T \mathbf{S} (\mathbf{C}_o)_{\text{frz}}. \tag{4.4}$$

With the 1PDM defined as in Eq. (4.3), the energy functional $E[\mathbf{P}]$ is then minimized while maintaining the fragment-block-diagonal structure of the MO coefficient matrix, i.e., the MOs are "absolutely localized" on each fragment. Such a variational optimization, which is called "SCF for molecular interaction" (SCF-MI) [301] for historical reasons, can be performed by solving locally projected SCF equations [233, 234, 301] (we refer the reader to Ref. 234 for more technical details) or by gradient-based algorithms. The resulting ALMOs, as shown in Fig. 4.1, are thus polarized in a fully self-consistent fashion within a properly antisymmetrized wavefunction. Such on-fragment orbital relaxations, on the other hand, do not give rise to charge flow between fragments under the Mulliken definition. Therefore, the "CT-free" state is fully defined by these polarized ALMOs, and the CT energy can be obtained by further computing its energy difference against the unconstrained SCF solution (Eq. (4.2)).

As for many other CT models based on orbital-space partitioning, the original ALMO definition of CT energy lacks a well-defined basis set limit. [103, 316] This is because when the size of the employed basis set grows, an AO basis function assigned to a certain fragment can easily overlap those on other fragments in 3-space such that the boundary between intra- and interfragment relaxations becomes ambiguous. To address this shortcoming, a new approach was proposed by some of us [117] to construct the polarization subspace of

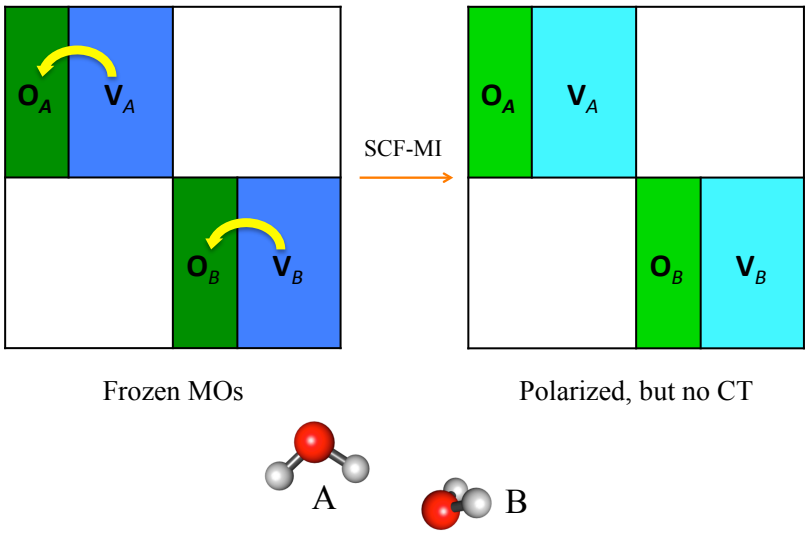

Figure 4.1: Illustration of the ALMO definition of the polarized yet "CT-free" state for the water dimer, obtained by performing a variational optimization (SCF-MI) from the frozen wavefunction. O and V stand for the occupied and virtual MOs on each fragment.

each fragment using their fragment electrical response functions (FERFs), whose basic spirit is to truncate the virtual space of each fragment so that degrees of freedom irrelevant to responses to external electric fields (polarization) are excluded in the SCF-MI calculation. The orbital response of a fragment to an external electric field can be captured by solving a set of coupled-perturbed (CP)SCF equations:

$$H_{ai,bj}(\Delta_\mu)_{bj} = (M_\mu)_{ai}, \tag{4.5}$$

where $\mathbf{H}$ is the SCF orbital Hessian and $\mathbf{M}_\mu$ is one component of the multipole matrix at a certain order, i, j and a, b are indices for occupied and virtual fragment MOs, respectively. The truncation of the original virtual space (spanned by $\mathbf{C}_v$) is achieved by performing a singular value decomposition (SVD) for each Δ_μ:

$$(\Delta_\mu)_{bj} = (L_\mu)_{ba}(d_\mu)_{ai}(R_\mu)_{ij} \tag{4.6}$$

$$\mathbf{V}_\mu = \mathbf{C}_v\tilde{\mathbf{L}}_\mu \tag{4.7}$$

where $\tilde{\mathbf{L}}_\mu$ refers to the first n_{occ} vectors of $\mathbf{L}_\mu$ (i.e. only n_{occ} virtual orbitals determine the exact linear response to $\mathbf{M}_\mu$).

Numerical results presented in Ref. 117 show that including both the three dipole- (D) and five quadrupole-type (Q) response functions per occupied orbital is sufficient to correctly reproduce the asymptotic behavior of electrical polarization, as well as providing a well-behaved basis set limit. The corresponding polarization subspace of a given fragment A

is

$$\begin{aligned}\mathrm{Pol}_A = \mathbf{O}_A \oplus \mathrm{span}\{\mathbf{V}_{Dx}, \mathbf{V}_{Dy}, \mathbf{V}_{Dz}\} \oplus \\ \mathrm{span}\{\mathbf{V}_{Q2,-2}, \mathbf{V}_{Q2,-1}, \mathbf{V}_{Q2,0}, \mathbf{V}_{Q2,1}, \mathbf{V}_{Q2,2}\}\end{aligned} \tag{4.8}$$

This model, with $8n_{\mathrm{occ}}$ virtual FERFs, is denoted as "FERF-nDQ" ("n" means that the MOs are *nonorthogonal* across fragments) and is employed in all FERF-related calculations in this paper.

4.2.2 CDFT-based definition of CT

The details of CDFT calculations have been documented in the previous papers by Wu and Van Voorhis,[293, 339] and here we briefly recapitulate the basics. The energy functional to be optimized can be represented as

$$E[\rho] = E_0[\rho] + \sum_c \lambda_c \left(\int \rho(\mathbf{r}) w_c(\mathbf{r}) \mathrm{d}\mathbf{r} - N_c \right) \tag{4.9}$$

where $E_0[\rho]$ is the standard Kohn-Sham (KS) energy functional. For each constraint in Eq. (4.9), $w_c(\mathbf{r})$ is the weighting function that corresponds to the constrained property, N_c is the value to be imposed in the calculation, and λ_c is the Lagrangian multiplier. The associated Fock matrix also contains an additional constraining potential:

$$\mathbf{F} = \mathbf{F}_0 + \sum_c \lambda_c \mathbf{W}_c, \tag{4.10}$$

where $\mathbf{F}_0 = \partial E_0 / \partial \mathbf{P}$, and $\mathbf{W}_c$ is the weighting function represented in the AO basis:

$$(W_c)_{\mu\nu} = \int w_c(\mathbf{r}) \chi_\mu(\mathbf{r}) \chi_\nu(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{4.11}$$

The energy functional given by Eq. (4.9) can thus be optimized through a double-loop SCF approach. While the outer loop updates $\mathbf{F}_0$, the inner loop searches for a set of Lagrangian multipliers to ensure that the density from diagonalizing $\mathbf{F}$ satisfies the constraints, using the first and second derivatives of E with respect to λ_c's whose forms have been previously derived. [293, 339]

CDFT has been successfully applied to the study of ET reactions that involve integer-number electron transfer from donor to acceptor, [346–349] while its application to modeling CT in the context of intermolecular interaction appeared more recently. The latter category of problems is more challenging for CDFT, because the result can be sensitive to the choice of partitioning schemes and the means used to generate constrained values (initial fragment charge populations). The DEDA by Wu et al. [113] employs the Becke-weight partitioning scheme [34] (using smoothly connected Voronoi cells): for each fragment, zero weights are assigned to grid points out of its corresponding Becke cell, and the fragment weight

matrix can thus be calculated using Eq. (4.11). The initial population of each fragment is determined by projecting the promolecule density (sum of isolated fragment densities) onto the corresponding weighting function. Using $\rho_0(\mathbf{r})$ to represent the promolecule density, the energy functional can be rewritten as

$$E[\rho] = E_0[\rho] + \sum_A \lambda_A \int (\rho(\mathbf{r}) - \rho_0(\mathbf{r}))\, w_A(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{4.12}$$

where the sum is over fragments whose populations are constrained.

The above-mentioned approach (which is denoted as "CDFT(Becke)" in this paper) was recently revisited by others [103, 342] and was found to yield reasonable CT stabilization energies. We should bear in mind that with this approach, the number of electrons to be constrained on each fragment ($N_A = \int \rho_0(\mathbf{r}) w_A(\mathbf{r}) \mathrm{d}\mathbf{r}$) is usually *not* an integer, which might be unnatural to think about. The option to impose integer fragment charges has also been investigated by Řezáč and de la Lande with various population schemes, [341] and it was demonstrated that a fragment-based Hirshfeld (FBH) partitioning can yield reasonable results while conserving integer fragment charge populations. The corresponding weighting functions are evaluated using isolated fragment densities ($\rho_A(\mathbf{r})$):

$$w_A^{\mathrm{FBH}}(\mathbf{r}) = \frac{\rho_A(\mathbf{r})}{\sum_A \rho_A(\mathbf{r})} = \rho_A(\mathbf{r})/\rho_0(\mathbf{r}) \tag{4.13}$$

It should be noted that the "Becke" scheme investigated in Ref. 341 also imposes integer charge population on each fragment so it behaves in a completely different way than the CDFT(Becke) approach introduced above.

In this work, we investigate both the Becke and FBH population schemes for CDFT calculations of CT. These real-space partitions ensure that the *net charge flow* between fragments is zero in the 3-space, which is illustrated in Fig. 4.2.

4.2.3 Characterization of constant-density CT

The frozen state in Wu's DEDA [113] is determined by optimizing the supersystem wavefunction subject to the constraint that the 3-space density is unchanged relative to the promolecule density, using a constrained search algorithm. [350] This constraint is much stronger than conserving fragment charge populations so the resulting variational energy is guaranteed to be an upper bound to that given by optimizing Eq. (4.12).

In order to characterize this constant-density relaxation effect, Horn and Head-Gordon proposed an alternative approach [351] employing a penalty function for the difference in 3-space density, where the metric of the error is chosen to be the Coulomb repulsion of the difference density with itself. The energy functional can be written as

$$E[\mathbf{P}] = E_0[\mathbf{P}] + \lambda \int \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \Delta\rho(\mathbf{r}_1) \frac{1}{r_{12}} \Delta\rho(\mathbf{r}_2) \tag{4.14}$$

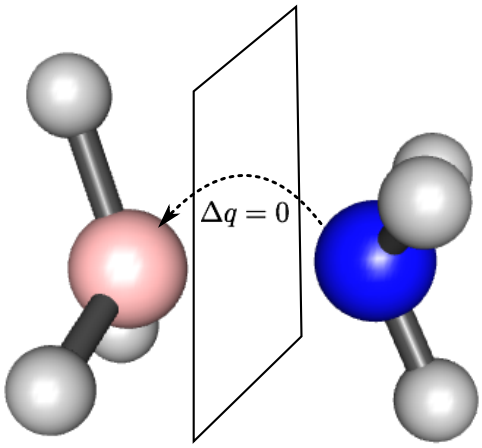

Figure 4.2: Illustration of the CDFT definition of the "CT-free" state. The plane stands for a partition of the 3-space, across which the net charge flow is constrained to be zero. In practice, this is accomplished by enforcing constant values of the total fragment charge, by measures such as the Becke and fragment-based Hirshfeld (FBH) definitions.

where $\Delta\rho(\mathbf{r}) = \rho(\mathbf{r}) - \rho_{\text{target}}(\mathbf{r})$, and λ is a scaling parameter that controls the magnitude of the penalty, which can be monotonically increased to impose the constraint as strictly as possible. It has been shown in Ref. 351 that when the target density is set to be the sum of isolated fragment densities, this approach is able to reproduce the energy lowering associated with the constant density relaxation in DEDA when λ is moderately large (note that both methods are unable to impose the density constraint exactly). In this paper, however, we use the density associated with the properly antisymmetrized frozen wavefunction (ρ_{frz}) as the target density exclusively.

With a given λ, either SCF-MI (enabling on-fragment relaxations only) or full SCF calculations could be performed to minimize the energy functional given by Eq. (4.14). The energy lowering associated with the former is defined as the constant-density (const-ρ) polarization, while the energy difference between the two corresponds to the const-ρ CT. Subject to a population scheme based on real-space partitioning, const-ρ CT occurs without changing fragment populations so it is not characterized as CT by CDFT-based models. Therefore, the evaluation of this quantity can provide useful insights into the numerical difference between ALMO and CDFT definitions, as illustrated in Sec. 4.3.

4.2.4 Characterization of the observable effects of CT

The methods introduced above are concerned with the evaluation of CT energy at a single geometry. In order to characterize the observable consequences of the components of intermolecular interactions, a reformulation of the ALMO-EDA in an adiabatic picture was recently reported by some of us. [119] In this scheme, the geometry of the intermolecular

complex is relaxed on the potential energy surface (PES) of each intermediate state (frozen, polarized, and fully relaxed) following the associated analytical gradients until a stationary point is reached. Molecular properties other than geometry, such as vibrational frequencies, multipole moments, can also be evaluated at these stationary points. By comparing the properties calculated on the "CT-free" state and those on the fully relaxed state, the effect of CT on these observables can be exhibited. Such calculations have also been reported by others using the (original) ALMO-EDA [119, 300] and the BLW-EDA [112, 298, 299, 352] , as they share the same definition for the "CT-free" state whose analytical nuclear gradient is easy to obtain. [303]

The equations required to perform an adiabatic ALMO-EDA are available in Ref. 119. In this work, we extend the concept of "adiabatic EDA" to the CDFT definition of CT, and the observable consequences of CT thus computed can be compared against those obtained by using the ALMO definition. The equations for the nuclear gradient of CDFT with the Becke partitioning scheme are presented in Appendix C.1.

4.3 Results

4.3.1 Computational details

The calculations in this work are performed with a development version of the Q-Chem 4.4 software package,[255] where the energy calculation using CDFT(FBH) and the nuclear gradient of the CDFT(Becke) approach were implemented for this work. Unless otherwise specified, the ωB97X-V [72] functional, which is a range-separated hybrid GGA with the VV10 [71] dispersion correction, is employed for the calculations. Based on extensive benchmarking [25, 80, 81] and our experience from previous EDA studies [119, 167, 169], this functional is amongst the most accurate available for intermolecular interactions. A fairly dense (75, 302) grid (75 radial shells for each atom with 302 Lebedev points in each) is employed for the integration of the exchange-correlation (XC) functional, while a smaller SG-1 grid [258] is used for the VV10 non-local correlation functional. The employed basis sets are specified with the results presented below. All the SCF calculations are converged to 10^{-8} a.u., and the geometry optimizations are converged when the maximum component of the nuclear gradient is below 10^{-4} a.u..

Unless otherwise noted, the CT stabilization energies based on the ALMO definition are evaluated with the aforementioned FERF-nDQ model, while the original ALMO approach based on AO-blocking is used for adiabatic EDA calculations, as the analytical gradient of SCF-MI using FERFs is not available yet. For the constant-density optimizations (see Sec. 4.2.3), the penalty parameter λ is chosen to be 2.0×10^3 as suggested in Ref. 351, and the FERF-nDQ model is employed to further separate the const-ρ energy lowering into POL and CT contributions.

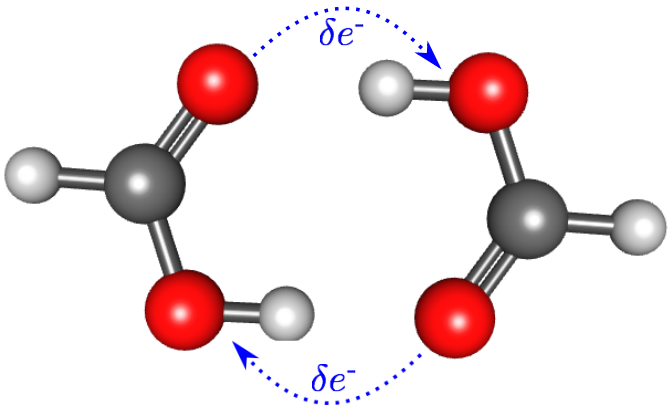

Figure 4.3: Geometry of the formic acid dimer (of C_{2h} symmetry). The arrows indicate that the net charge flow between two formic acid molecules is zero, even though CT will occur between the monomers in each of the two identical hydrogen bonds.

4.3.2 Cyclic H-bonding system

The difference between the ALMO and CDFT definition of CT can be most clearly revealed via a cyclic, double hydrogen-bonded system, such as the formic acid dimer (Fig. 4.3), as well as the *p*-biphthalate dimer (a so-called "anti-electrostatic H-bonding" (AEHB) complex) studied by us in a previous paper [97], and the pyridine dimer (C_{2h}) studied in Ref. 102. To simplify the discussion, below we focus on the formic acid dimer.

The CT energies at equilibrium geometry and CT-induced changes in the relevant observables calculated by original ALMO, FERF-nDQ and CDFT are presented in Table. 4.2 (the original data for the observable calculations are available in Table C.1). The most noticeable feature is that the energy contribution from CT is zero with the CDFT definition, and therefore so too is the effect of CT on the observables. This can be easily understood through the formalism of the CDFT approach: due to the symmetry of this system, the forward and backward donations across the molecular boundary (under a given definition) exactly cancel each other in terms of the resulting population change, which thus automatically satisfies the constraint on fragment populations. As a result, the "CT-free" state becomes identical to the fully relaxed state under the CDFT definition.

The ALMO-based methods, on the other hand, suggest a significant contribution from CT for this system. CT accounts for roughly 50% of the equilibrium interaction energy (-82.27 kJ/mol), which is a usually larger proportion than in other typical neutral H-bonding complexes (see Sec. 4.3.3). The calculated CT-induced changes in the proton-donating O–H bond, such as its elongation and the red shift in its stretch frequency, are also remarkable. As argued in Ref. 117, the ALMO-based definition of the "CT-free" state not only conserves fragment populations (under the Mulliken definition), it also suppresses charge flow between fragments provided that an appropriate polarization subspace (e.g. the one given by Eq. (4.8))

Table 4.2: Equilibrium CT stabilization energies (in kJ/mol) and CT-induced changes in observables for the formic acid dimer calculated with three distinct models. Δr_{OH} and $\Delta r_{\mathrm{O\cdot\cdot H}}$ (in Å) refer to the changes in the lengths of the proton-donating O–H bond and the hydrogen bond (O··H distance), respectively. $\Delta\angle_{\mathrm{OH\cdot\cdot O}}$ (°) is for the increase in the hydrogen-bond angle due to CT, and $\Delta\nu_{\mathrm{OH}}$ for the red shift (in cm^{-1}) in the OH stretch frequency. All the calculations are performed at the ωB97X-V/def2-TZVPPD level of theory.

	ALMO	FERF	CDFT
ΔE_{CT}	-45.92	-48.94	0
Δr_{OH}	0.024	0.025	0
$\Delta r_{\mathrm{O\cdot\cdot H}}$	-0.33	-0.35	0
$\Delta\angle_{\mathrm{OH\cdot\cdot O}}$	5.26	5.28	0
$\Delta\nu_{\mathrm{OH}}$	468	478	0

is chosen for each monomer. Therefore, many orbital rotation degrees of freedom allowed in CDFT are forbidden in an ALMO-based calculation. It is also noteworthy that with a moderately large basis set (def2-TZVPPD [183]), the results given by AO-based ALMO and FERF are very similar for this system, which, to some extent, validates the use of the original ALMO model for molecular property calculations in this work.

From this example we see the first important difference between the ALMO and CDFT definitions of CT: the CDFT definition relies on the existence of net population change, and the contributions from forward and backward donations therefore cancel each other, which is different than the ALMO definition (and other definitions using orbital-space partitioning) in which forward and backward donations contribute in an additive manner. In cases like this the CDFT constraint defining the "CT-free" state is demonstrably too weak, and that state is in fact contaminated with 100% of the energy lowering due to CT.

4.3.3 Unidirectional H-bonding systems

The next set of H-bonding complexes (X–H$\cdots$Y) we investigate is presumably dominated by donation in one direction (Y to X–H) in terms of the CT effect. Therefore, the mutual cancellation of forward and backward contributions in CDFT should be a less pronounced issue here. The ALMO-EDA results, as well as CT energies calculated by CDFT (with both Becke and FBH population schemes) and energy lowerings due to constant-density relaxation, are collected in Table 4.3. The total binding energies for the first two complexes are similar to each other, and the contributions from CT, based on the ALMO definition, are also close to each other, accounting for roughly one third of the total binding energy. A similar proportion of CT holds for the water-Cl^- complex, despite the larger magnitude of both the total interaction energy and the CT component due to the existence of an anion.

A rather different picture is given by the CDFT-based approaches. With the CDFT(Becke) method, the resulting CT energies are 67%, 42%, and 87% smaller than

Table 4.3: Energetic results (in kJ/mol) of ALMO-EDA (using FERFs), CDFT CT, and constant-density (const-ρ) relaxation for three H-bonding systems. All calculations are performed with ωB97X-V/def2-TZVPPD at the equilibrium geometries optimized at the same level of theory.

	H_2O dimer	HF dimer	water-Cl^-
FRZ	-9.40	-7.38	-29.99
POL	-4.59	-5.32	-15.32
CT	-7.16	-6.93	-19.83
Total	-21.15	-19.63	-65.14
CT (CDFT, Becke)	-2.39	-4.03	-2.51
CT (CDFT, FBH)	-3.54	-4.39	-7.85
const-ρ POL	-0.45	-0.30	-0.82
const-ρ CT	-5.05	-4.39	-13.77

the corresponding ALMO results for these systems, respectively. The FBH partitioning scheme yields larger CT energies for this set of systems, especially for the water-Cl^- complex. However, there is still a significant gap compared to the results computed with the ALMO definition.

In order to shed some light on this difference (i.e., is the CDFT constraint defining the "CT-free" state again too weak?), the effect of constant-density relaxation is characterized by performing penalized SCF-MI and SCF calculations as introduced in Sec. 4.2.3. The energy lowerings are measured against the energy of the frozen wavefunction, and the 3-space density is constrained to be as close to ρ_{frz} as possible. While the effect of const-ρ POL is relatively small, the magnitude of the constant-density interfragment relaxation (CT) suggested by these calculations is striking, although we must note that the imposition of the const-ρ constraint is inexact. For the water dimer and the water-Cl^- complex, the value of const-ρ CT is rather comparable to the gap between ALMO and CDFT CT energies (the energetics for the water dimer is further illustrated in Fig. 4.4). This finding demonstrates how nature enables intermolecular orbital relaxation without altering the charge population on each fragment (and even the 3-space electron density). Such constant-density intermolecular relaxation is not incorporated in the CDFT definition of CT, as it does not violate the constraint imposed by real-space partitions and thus it already occurs on the "CT-free" state, which explains why CDFT-based approaches usually yield smaller CT energies than orbital-based methods. It is remarkable that in some cases (such as the HF dimer), the energy lowering due to const-ρ CT is even larger than the difference between the ALMO and CDFT results, which, however, is still possible as long as the system energy after const-ρ relaxation is no lower than that of the CDFT state.

The elongation of the X–H bond and the red shift in its stretch frequency are well-accepted fingerprints of CT in typical hydrogen-bonding systems. By using the adiabatic EDA, we calculated the CT-induced changes in these two observables, and the results are

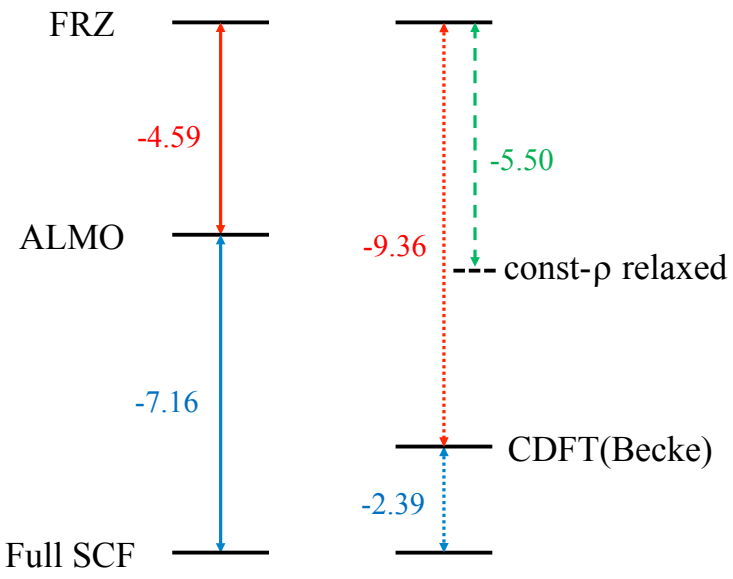

Figure 4.4: Illustration of the energetics as calculated in Table 4.3 for the water dimer system. Each horizontal line refers to an intermediate state, and each vertical double-headed arrow corresponds to one type of relaxation, whose associated energy lowering is indicated by the number beside it (in kJ/mol). The dashed horizontal line that is obtained by performing const-ρ relaxation should be close to the energy of the frozen state in DEDA.

shown in Table 4.4. For the H_2O dimer and the HF dimer, the observable changes calculated by ALMO and CDFT are not remarkably different, and they can be characterized as an effect of CT based on the result of either approach. The differences are also consistent with the trend CDFT constraint being slightly weaker than the ALMO constraint. However, qualitatively different pictures are generated ALMO and CDFT for the water-Cl^- complex, which is more clearly illustrated in Fig. 4.5. We note that an adiabatic ALMO-EDA study of this system has been reported in a previous paper by us (Ref. 119), while here we further include the result calculated at the stationary point on the CDFT surface. On the ALMO PES, the optimal structure is roughly midway between the frozen and fully relaxed structures in terms of the $Cl\cdots H_d$ distance and the $Cl\cdot\cdot H_d$–O angle (H_d refers to the donated proton), with a slightly lengthened O–H_d bond compared to that on the frozen PES. It was also revealed that the energy of this structure is extremely close to that of another stationary geometry in which Cl^- lies on the bisector of the H_2O molecule, [119] which further suggests that the near-linear hydrogen bond in the final complex is an effect of CT. The structure optimized on the CDFT PES, in contrast, is rather close to the fully relaxed one, with the O–H_d bond already significantly elongated by 0.02 Å. The red shift in the OH_d stretch frequency, which can be viewed from its difference against another OH stretch involving the free proton (denoted as $\Delta\nu$ in Fig. 4.5), is also much more significant on the CDFT PES compared to the value of $\Delta\nu$ calculated on the ALMO surface. As these fingerprints

Table 4.4: X–H bond lengths (Å) and XH stretch frequencies (cm^{-1}) for H-bonding systems X–H$\cdots$Y calculated on the frozen, ALMO, CDFT(Becke), and fully relaxed PES using ωB97X-V/def2-TZVPPD. In order to decouple the two FH modes in the HF dimer, the non-H-bonded hydrogen is substituted with deuterium. For the frozen structure of $H_2O{\cdot\cdot}Cl^-$, the lower (symmetric) OH stretch frequency is reported.

	H_2O dimer		HF dimer		$H_2O{\cdot\cdot}Cl^-$	
	r_{OH}	ν_{OH}	r_{FH}	ν_{FH}	r_{OH}	ν_{OH}
FRZ	0.961	3861	0.922	4120	0.962	3866
ALMO	0.962	3853	0.923	4095	0.968	3803
CDFT(Becke)	0.964	3829	0.924	4086	0.981	3551
Full	0.967	3754	0.928	3981	0.987	3406

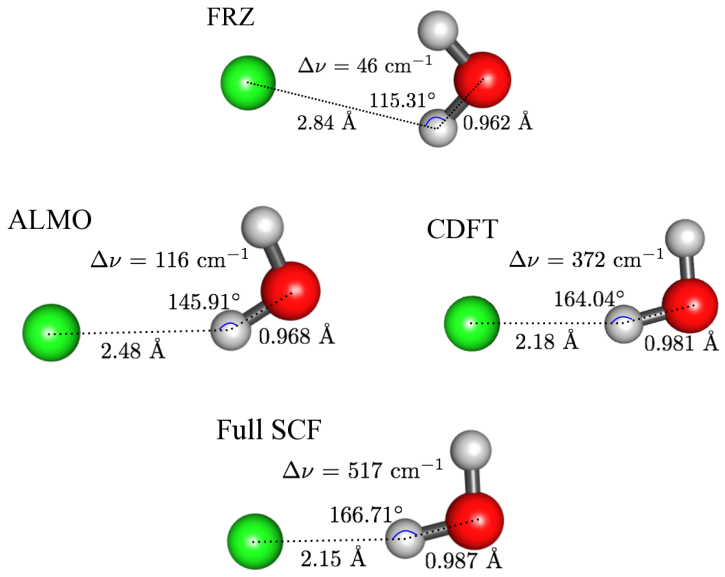

Figure 4.5: Adiabatic EDA results for the $H_2O\cdots Cl^-$ complex computed with ωB97X-V/def2-TZVPPD. Changes in the Cl$\cdots$H$_d$ distance, the Cl$\cdot\cdot$H$_d$–O angle, length of the O–H$_d$ bond, and the split between the two OH stretch modes in the H_2O molecule ($\Delta\nu$) are indicated in the figure. Two definitions of the polarized yet "CT-free" state (ALMO and CDFT(Becke)) are compared.

of CT already appear on the PES that is supposed to be "CT-free", the CDFT approach thus suggests a minimal effect of CT on the observable shifts in the water-Cl^- complex. Or, stated another way, the "CT-free" CDFT state is in fact significantly CT-contaminated, akin to formic acid dimer case.

Table 4.5: ALMO-CTA results for NH_3-BH_3 and BH_3-CO complexes calculated with ωB97X-V/def2-TZVPP. The energy changes (ΔE) are in kJ/mol, and the charge delocalizations (ΔQ) are in me^-. For both complexes, "D→A" refers to the donation towards BH_3, and "A→D" is for the opposite direction. The high-order (HO) contribution stands for the difference between CT energies calculated by performing a single Roothaan step (RS) and by converging the full SCF solution, which is not further decomposed into D→A and A→D contributions.

	D→A		A→D		ΔE_{HO}
	ΔE	ΔQ	ΔE	ΔQ	
NH_3-BH_3	-106.60	34.75	-9.58	4.02	-24.14
BH_3-CO	-85.83	22.02	-60.22	41.87	-45.30

4.3.4 Borane compounds

Borane (BH_3), as a typical Lewis acid because of its electron-deficient property, can form donor-acceptor complexes with Lewis bases (electron-donating species). The ammonia-borane (NH_3-BH_3) and the borane-carbonyl (BH_3-CO) complexes are two textbook examples for Lewis acid-base adducts, and they have been investigated by various approaches such as the first-generation ALMO-CTA [305] and the regularized SAPT method. [102] Here we revisit these two systems to further explore the differences between the ALMO and CDFT definitions. According to the ALMO-CTA results calculated with ωB97X-V and the slightly smaller def2-TZVPP [353] basis (Table 4.5), the CT effect in NH_3-BH_3 is dominated by the forward donation towards BH_3, while in the case of BH_3-CO, there is a significant contribution from the back-donation (BH_3 →CO), which was characterized as from the B–H σ bonds to the π^* anti-bonding orbital of CO by an analysis using the complementary occupied-virtual pairs (COVPs). [305] Therefore, we expect that the CT stabilization energy given by CDFT is small for the BH_3-CO complex due to the mutual cancellation of forward and backward donations under this definition.

The results in Table 4.6 are consistent with our expectation. For the BH_3-CO complex, the CT energy given by CDFT is close to zero with either population scheme, indicating the near-perfect cancellation of the σ donation of CO with the $\sigma \to \pi^*$ back donation from BH_3 in terms of the resulting population change in the 3-space. The ALMO-based definition, on the other hand, suggests a substantial contribution from CT, as the forward and backward donations contribute to the total ALMO CT energy in an additive manner. Unlike BH_3-CO, the CT energies given by CDFT gain much larger magnitude for the NH_3-BH_3 complex due to the unidirectionality of its electron donation, which, however, is only about 1/3 as large as the ALMO CT energy. Note that the CT energy for NH_3-BH_3 calculated by CDFT(FBH) is in good agreement with the value reported in Ref. 341 using the same population scheme (-11.99 kcal/mol, calculated by PBE/def2-QZVPPD). This large difference between ALMO and CDFT CT energies, again, is fairly comparable to the magnitude of const-ρ CT: if we use the average of the Becke and FBH results to represent the CDFT value, then const-ρ CT

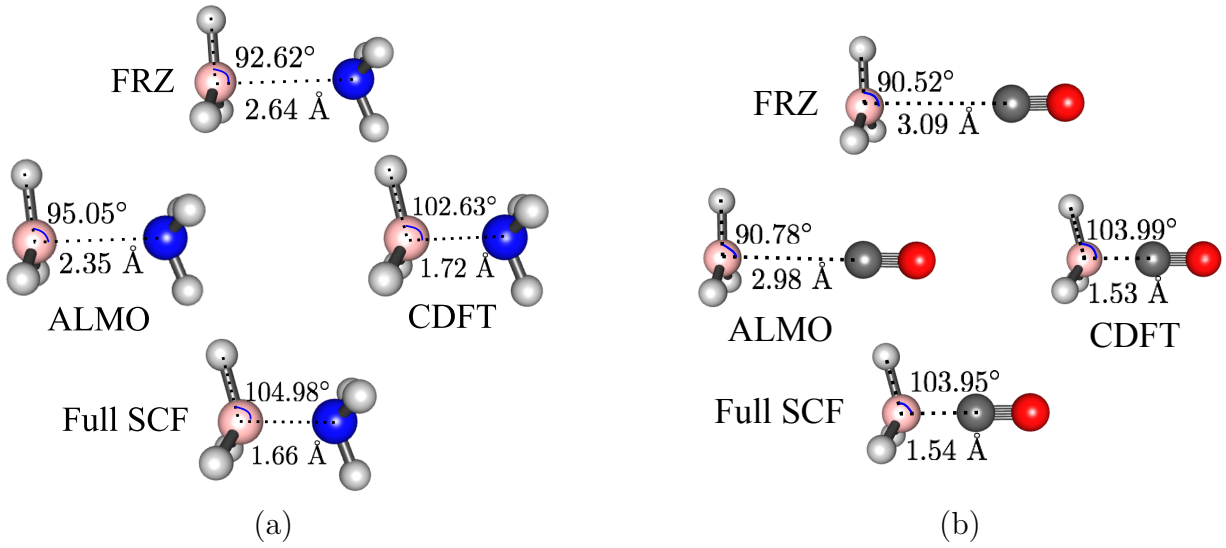

Figure 4.6: Adiabatic EDA results for the (a) NH_3-BH_3 and (b) BH_3-CO complexes calculated with ωB97X-V/def2-TZVPP. The B$\cdots$Y distance and $\angle$Y$\cdot\cdot$B–H at each equilibrium structure are denoted (Y refers to the atom on the Lewis base that the donating lone pair is associated with). Two definitions of the polarized yet "CT-free" state (ALMO and CDFT(Becke)) are compared.

Table 4.6: Energetic results (in kJ/mol) of ALMO-EDA (using FERFs), CDFT CT, and constant-density (const-ρ) relaxation for the NH_3-BH_3 and BH_3-CO complexes. All calculations are performed with ωB97X-V/def2-TZVPPD at the equilibrium geometries optimized at the same level of theory.

	NH_3-BH_3	BH_3-CO
FRZ	117.29	302.13
POL	-150.91	-187.20
CT	-155.88	-278.81
Total	-189.50	-163.88
CT (CDFT, Becke)	-63.62	-0.64
CT (CDFT, FBH)	-52.63	-0.56
const-ρ POL	-20.09	-34.19
const ρ CT	-85.23	-153.84

accounts for 87% of the gap between ALMO and CDFT results. In this sense, the NH_3-BH_3 complex is similar to the unidirectional H-bonding systems discussed in Sec. 4.3.3. Despite the even larger const-ρ CT in BH_3-CO, it is far from addressing the difference between the ALMO and CDFT results for this system, implying that the above-mentioned cancellation of bidirectional donations be the more important reason.

It was pointed out in Ref. 119 that conventional ALMO-EDA calculations performed

at the fully relaxed geometry of a strong donor-acceptor complex often result in a strongly repulsive FRZ term and an overemphasized role of POL, due to the small intermolecular distance driven by the significant CT effect. This seems to be the case for the results in Table 4.6, and thus it is useful to revisit these systems with the adiabatic EDA. While the full adiabatic EDA results for NH_3-BH_3 and BH_3-CO (including the energetics) are available in Table C.2, the key features are exhibited in Fig. 4.6. Since the AO-based ALMO definition is used in these calculations, the def2-TZVPP basis which contains no diffuse functions is employed. Based on the ALMO definition, the significance of polarization is less pronounced under the adiabatic picture, and the key geometric features of these donor-acceptor complexes, including the bending of the originally planar BH_3 molecule (the hybridization of the boron atom changes from sp^2 to sp^3) and the short intermolecular distance, are mainly attributed to CT. These structural features, however, already appear in the CDFT-optimized geometries for both systems. It is clearly revealed in Fig. 4.6 that while the structures optimized on the ALMO PES are closer to the frozen structures, those optimized with CDFT are more similar to the fully relaxed structures. For BH_3-CO, the CDFT-optimized structure even "overshoots" the fully relaxed one by having a slightly shorter B$\cdots$C distance and a more bent BH_3 plane. Although this does not go against any basic principles in adiabatic EDA, it is a peculiar result especially for a strong donor-acceptor complex like BH_3-CO. We think that it is most likely related to the specific partitioning scheme (Becke) employed here: the equilibrium geometry optimized with CDFT(FBH) (using finite-difference gradient) has r(B$\cdot\cdot$C) = 1.55 Å and $\angle$C$\cdot\cdot$B–H = 103.77°, which look more reasonable in comparison.

4.3.5 Metal-carbonyl complexes

Metal-carbonyl complexes are another important class of donor-acceptor adducts whose charge-transfer effects have been previously studied by approaches such as CSOV, [354] BLW-EDA,[112] and ALMO-CTA.[305] Based on the change in CO stretch frequency upon the complexation, they can be categorized into "classical" (whose ν(CO) is red-shifted) and "nonclassical" (ν(CO) blue-shifted) metal-carbonyl complexes.[280, 355] According to the ALMO- or BLW-based analyses, [112, 305] the classification is determined by the relative strength of the backward donation (M(d)$\rightarrow$CO(π^*)) and the forward donation that forms the σ(M–C) bond.

Here we choose two simple monocarbonyl complexes $Cu(CO)^+$ and Ni(CO) that were previously investigated by Mo et al. [112] to show the difference in ALMO and CDFT results, which, according to Table 7 in Ref. 112, are representative of the above-mentioned "nonclassical" and "classical" cases, respectively. In order to verify that our previous findings still hold upon a change of functional, and to compare with the results in Ref. 112, the B3LYP functional [42] with a Becke-Johnson damped D3 correction [58] is employed for this set of calculations. The energetic results computed at the equilibrium structures are summarized in Table 4.7. The ALMO-EDA results suggest that POL is the most significant binding force for the $Cu(CO)^+$ complex, while Ni(CO) is more similar to the borane complexes (see Table 4.6) whose FRZ and POL both have very large magnitude but are of opposite signs (due

Table 4.7: ALMO-EDA (using FERFs) and CDFT results (in kJ/mol) for the $Cu(CO)^+$ and Ni(CO) model complexes. All calculations are performed with B3LYP-D3(BJ)/def2-TZVPPD at the equilibrium geometries optimized at the same level of theory. The spin state of the Ni atom remains singlet throughout, i.e., the energy consumed for the spin-state promotion is not included.

	$Cu(CO)^+$	Ni(CO)
FRZ	12.75	171.41
POL	-128.50	-220.11
CT	-48.60	-232.97
CT (CDFT, Becke)	-8.54	-32.79
CT (CDFT, FBH)	-5.62	-37.69
INT	-164.35	-281.67

Table 4.8: Summary of adiabatic EDA results for $Cu(CO)^+$ and Ni(CO) computed with B3LYP-D3(BJ)/def2-TZVPP. The distances are in Å and the vibrational frequencies are in cm^{-1}. The Ni(CO) complex remains to be singlet at all stages.

	$Cu(CO)^+$			Ni(CO)		
	R(M–C)	R(C–O)	ν(CO)	R(M–C)	R(C–O)	ν(CO)
FRZ	2.328	1.112	2338	2.831	1.123	2231
ALMO	2.027	1.111	2362	1.923	1.117	2321
CDFT	1.858	1.116	2320	1.729	1.136	2192
Full SCF	1.905	1.114	2314	1.678	1.148	2090
Free CO		1.125	2215			

to the close intermolecular contact). The CT component, nevertheless, still has the largest contribution to the latter system . Note that the ALMO-EDA results here are remarkably different from the BLW-EDA results presented in Ref. 112: the CT energies reported in the latter are much larger for both systems, and the POL components are correspondingly smaller. The reason for this difference is unclear to us, as the FERF model employed in our calculations, by construction, should yield even larger CT energies than those obtained by using the original ALMO model, where the latter is formulated in the same way as the BLW approach.

The CDFT-based CT energies, on the other hand, are roughly one order of magnitude smaller than the ALMO results for both systems, although the relative strength of CT between these two complexes is correctly reflected. This is most likely because the forward (σ) and backward ($d \rightarrow \pi^*$) donations are both not negligible in these complexes, as suggested in Refs. 112 and 305, and they cancel each other under the CDFT definition.

The significance of the M→CO backward donation can be revealed from the shift in CO stretch frequency (ν(CO)) relative to that for an isolated CO molecule. Here we utilize

the adiabatic EDA approach again to track the changes in ν(CO) and the M–C and C–O distances, and the results are summarized in Table 4.8. On the frozen PES, ν(CO) is shifted to a higher frequency and the length of the C–O bond is reduced. These changes are substantial for the cationic $Cu(CO)^+$ complex while being much smaller for the neutral Ni(CO) system whose optimal M–C distance on the frozen PES is also over 0.5 Å larger. We note that the blue shift in CO stretch frequency on the frozen surface is an interesting finding provided by the adiabatic ALMO-EDA, and it might be related to other phenomena such as blue-shifting hydrogen bonds, [352, 356, 357] although a detailed analysis of this is obviously beyond the scope of the present paper. Polarization, as suggested by the ALMO model, further shortens the C–O bond and gives rise to an increased blue shift. The final ν(CO) is thus determined by CT: for the "classical" complex Ni(CO), ν(CO) is lowered by ∼230 cm^{-1} once CT is enabled, indicating a significant M(d)→CO(π^*) donation and rendering an eventually red-shifted ν(CO) (by 125 cm^{-1}) and a lengthened C–O bond (by over 0.02 Å); for the "nonclassical" complex $Cu(CO)^+$, the effect of CT is much smaller and ν(CO) is still blue-shifted by roughly 100 cm^{-1} on the fully relaxed surface.

In contrast, on the "CT-free" PES defined based on CDFT, ν(CO) is already lower than the frequency of free CO stretch and the C–O bond is also elongated, suggesting a rather different role of CT in modulating the properties of Ni(CO). The CDFT results for $Cu(CO)^+$ are even more extraordinary, as the resulting optimized geometry "overshoots" the fully relaxed structure by having a shorter R(Cu–C) and a longer C–O bond. The resulting ν(CO) is very close to the full SCF result as well. These results indicate that the M→CO donation already occurs on the CDFT surface, which thus seems to be an inappropriate definition for the "CT-free" state at least for the metal-carbonyl complexes.

4.3.6 Complexes of water and metal cations

The last set of systems that we investigate in this section comprises five complexes formed by water and alkali (Li, Na, K) or alkali earth (Mg, Ca) metal cations. A separation of POL and CT for these interactions is meaningful for the development of polarizable force fields, as shown in several works by us [167, 358] and others. [309, 310, 359] Using the second-generation ALMO-EDA, it was found that the magnitude of the equilibrium CT energies for these systems is ordered as Na < K < Li and Mg < Ca, [167] and the latter was further manifested through a study using the adiabatic ALMO-EDA. [119] On the other hand, it was reported in Ref. 103 that the CDFT(Becke) approach gives a different order (K < Na < Li) for the monovalent cations, which was considered to be more reasonable for these "obvious" cases as ΔE_{CT} was supposed to be larger with a shorter $O\cdots M^+$ distance.

The equilibrium CT energies for these systems evaluated with four distinct models (original AO-based ALMO, FERF-nDQ, CDFT(Becke), and CDFT(FBH)) are plotted in Fig. 4.7 (the original data are available in Table C.3). The ωB97X-D3 functional [360] is employed here in order to compare with the results in Ref. 103. Since a quadruple-ζ basis set (def2-QZVPPD) is used, there is a notable gap between the results of AO-based ALMO (which was used in Ref. 103) and FERF, especially for the complexes of smaller intermolecular

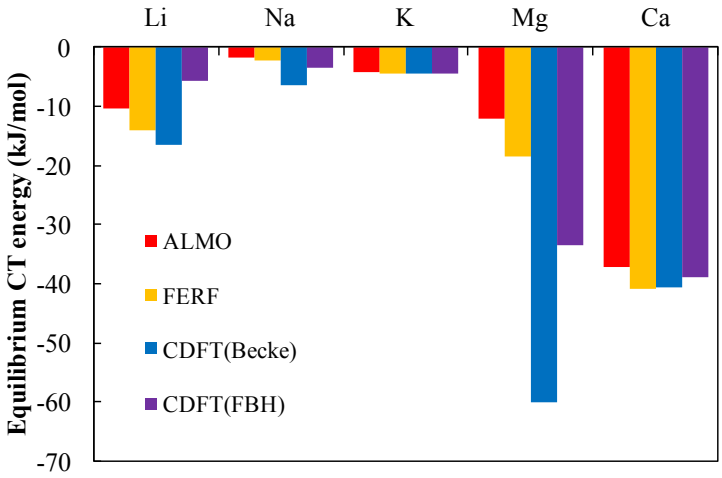

Figure 4.7: Equilibrium CT stabilization energies (in kJ/mol) for the water-M^{n+} (M^{n+}= Li^+, Na^+, K^+, Mg^{2+} and Ca^{2+}) complexes evaluated using AO-based ALMO (red), FERF (yellow), CDFT(Becke) (blue) and CDFT(FBH) (purple) methods. All the calculations are performed at the ωB97X-D3/def2-QZVPPD level of theory, while the equilibrium structures optimized with ωB97X-V/def2-QZVPPD are from a previous work. [167]

distances (water-Li^+, Mg^{2+}). Nonetheless, as opposed to all other systems discussed above, the CT energies given by FERF are still smaller than the CDFT(Becke) results for the complexes with Li^+, Na^+, and Mg^{2+}, while the gap becomes much smaller on K^+ and Ca^{2+} for which the FERF results are marginally more favorable. Fig. 4.7 also indicates the CDFT results are sensitive to the employed population scheme, as ΔE_{CT} is significantly decreased for the complexes with Li^+ (65%), Na^+ (46%), and Mg^{2+} (44%) once the Becke partitioning scheme is replaced by FBH (the percent values in the parentheses are reductions relative to the CDFT(Becke) values). Interestingly, the CT energies generated by CDFT(FBH) are ordered in the same way as the FERF results, despite the substantial energy differences between these two approaches for $H_2O\cdots Li^+$ and $H_2O\cdots Mg^{2+}$.

While there is no criterion for a unique correct order for the magnitude of CT energies in these systems, it is possible to shed some light on the strikingly large CT energies given by CDFT(Becke) for the water-Mg^{2+} and, to a lesser extent, water-Li^+ complexes. The Becke scheme essentially relies on a partition based on atomic Voronoi cells to obtain the value of the CDFT weighting function on each grid point, i.e., the 3-space density evaluated on a grid point is assigned to a fragment based on the shortest-distance criterion. Therefore, for systems whose $O\cdots M^{n+}$ distance is relatively small, such a partitioning scheme that completely neglects the electronic structure might place the interfragment boundary in a region where the electron density of water is not yet insignificant. As a consequence, the polarization of H_2O, which also pulls its electron density towards the cation, might be misinterpreted as CT across the boundary and thus be incorrectly inhibited in the CDFT(Becke) calculations. Table 4.9 shows the electron population on H_2O to be constrained based on the

Table 4.9: Number of electrons on the H_2O molecule that are constrained in the CDFT(Becke) calculations for the $H_2O\cdots M^{n+}$ complexes (using ωB97X-D3/def2-QZVPPD). The values are determined by projecting the promolecule 1PDM onto the Becke weight matrix associated with the H_2O fragment.

	Li^+	Na^+	K^+	Mg^{2+}	Ca^{2+}
Population (e^-)	9.84	9.95	10.06	9.88	10.03

Becke partition, which verifies our assumption above. For the complexes with Li^+, Na^+ and Mg^{2+}, the electron population on the H_2O fragment is less than 10, indicating that the Becke scheme divides the system such that a small portion of electron density of H_2O is assigned to the metal cation, which, as we argued above, might lead to an insufficiently polarized CDFT solution. On the other hand, the partitions in water-K^+ and water-Ca^{2+} allow them to be fully polarized in the CDFT calculation, and thus the CDFT(Becke) scheme does not yield excessively large CT energies for these systems. In this sense, FBH seems to be a more advantageous partitioning scheme as it makes use of the electronic structure information in the construction of the weight matrix.

Although there is no doubt that the strength of CT is closely related to parameters such as the intermolecular distance, the IP of the donor and the EA of the acceptor, one should not try to reach a conclusion only based on a subset of them. For instance, it was argued in Ref. 359 that the CT energies for divalent cations should have an order $Mg^{2+} > Ca^{2+}$ as the EA of the former is larger, which, at the same time, also matches the distance-based argument in Ref. 103. However, according to Fig. 4.7, such an order is only supported by the CDFT(Becke) scheme. Using the original ALMO-CTA, we plotted the most significant COVPs for the $H_2O\rightarrow Mg^{2+}$ and $H_2O\rightarrow Ca^{2+}$ donations in Fig. 4.8. Qualitative difference exists in the donor-acceptor orbital picture for these two systems: for $H_2O\cdots Mg^{2+}$, there are two significant COVPs and the acceptor orbitals on Mg^{2+} are an *sp*-hybridized orbital (in panel a) and an empty *p* orbital (in panel b), respectively; for the complex with Ca^{2+}, there is only one significant COVP (panel c) and the acceptor orbital is clearly an empty *d* orbital of Ca^{2+}. While the COVP analysis does not suffice to rationalize the larger CT energy in $H_2O\cdots Ca^{2+}$ than in $H_2O\cdots Mg^{2+}$, it implies that a simple argument based on the $O\cdots M^{n+}$ distance or the EA of M^{n+} solely may overlook some of the underlying details about CT.

4.3.7 Comparison with the regularized SAPT method

Before concluding this paper, we want to briefly discuss the recently proposed regularized SAPT method, [102] which provides a distinct perspective on CT. In this scheme, CT is considered to be a tunneling process induced by the Coulomb potential of fragment nuclei, as it is not fully screened by the electrons and becomes singular when $r \rightarrow 0$. Therefore, by calculating the second-order SAPT(DFT) induction energy with a regularized Coulomb

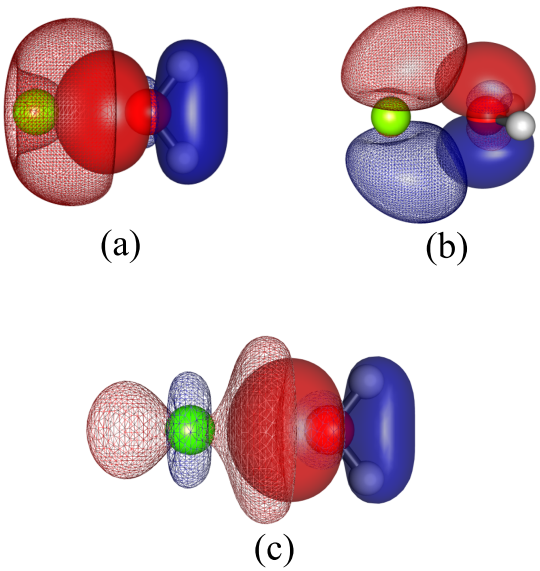

Figure 4.8: Dominating COVPs for the forward donations ($H_2O \rightarrow M^{n+}$) in the water-Mg^{2+} (a, b) and water-Ca^{2+} (c) complexes generated by the original ALMO-CTA using the smaller def2-SVPD basis. The donor (solid) and acceptor (meshed) orbitals are plotted with an isosurface value 0.05 a.u.

Table 4.10: Comparison of CT energies (in kJ/mol) evaluated with the regularized SAPT, ALMO (using FERFs) and CDFT approaches. The regularized SAPT results are taken from Table 1 in Ref. 102, and the ALMO and CDFT calculations are performed at the ωB97X-V/def2-TZVPPD level of theory. For consistency, geometries provided in the Supporting Information of Ref. 102 are used for the ALMO and CDFT calculations. The column "IND(total)" shows the sum of D→A and A→D contributions to the second-order induction energy.

	Regularized SAPT					ALMO		CDFT	
	CT(D→A)	CT(A→D)	IND(D→A)	IND(A→D)	IND(Total)	CT	POL+CT	CT(Becke)	CT(FBH)
H_2O dimer	-1.39	-0.10	-4.59	-1.25	-5.84	-5.53	-9.35	-2.02	-2.89
HF dimer	-1.33	-0.04	-5.90	-0.37	-6.27	-6.82	-12.06	-3.91	-4.27
FH···CO	-1.39	-0.05	-6.24	-0.23	-6.47	-9.26	-13.38	-3.21	-3.10
FH···OC	-0.37	-0.02	-3.04	-0.02	-3.06	-2.43	-5.29	-1.91	-1.60
Pyridine dimer	-0.18	-0.18	-1.63	-1.63	-3.26	-3.34	-6.63	0.00	0.00
NH_3-BH_3	-61.65	-9.84	-160.77	-14.83	-175.60	-142.97	-277.44	-64.79	-55.58
BH_3-CO	-139.16	-31.85	-296.24	-50.45	-346.69	-284.45	-477.64	-0.33	-1.13

operator that is screened in the short range ($1/r \rightarrow (1 - e^{-\eta r})/r$), a "CT-free" induction energy can be obtained. The CT energy can thus be computed in a subtractive way, as shown by Eq. (4.1). While the formulation of this method is completely different than the ALMO-based definition, they should at least have two common features: (i) the forward and backward contributions are constructive rather than canceling each other, and (ii) the definition of CT does not rely on net charge flow.

With the same set of systems as in Table 1 of Ref. 102, we compare the equilibrium CT energies evaluated with the regularized SAPT, ALMO, and CDFT definitions in Table 4.10. We note that for systems that were investigated above, the ALMO and CDFT results are slightly different here, as the geometries are directly taken from Ref. 102. The contrast between the ALMO and CDFT results is consistent with the trends that have been elucidated via the examples above, and it is notable that the pyridine dimer (C_{2h}), as another cyclic H-bonding complex, has zero CT according to the CDFT definition, which is clearly not the case for regularized SAPT.

However, for all complexes investigated in Table 4.10, the regularized SAPT method yields a smaller CT energy compared to the corresponding ALMO result; and for the unidirectional H-bonding systems (the first four complexes), its magnitude is even smaller than that of the corresponding CDFT result. We think that the relatively small magnitude of regularized SAPT CT energies might arise from the truncation of SAPT induction energy at the second-order: for each complex in Table 4.10, the total pairwise additive second-order induction energy, IND(Total), is significantly smaller than the sum of POL and CT suggested by ALMO-EDA. Recall that within the SAPT(DFT) theory, a "δHF" term (defined as the difference between the counterpoise-corrected Hartree-Fock interaction energy and the second-order SAPT energy excluding the dispersion and exchange-dispersion terms) is utilized to capture the missing high-order induction effect. With this term taken into account, the total SAPT(DFT) induction energy for the equilibrium water dimer, the first system in Table 4.10, was reported to be -2.2840 kcal/mol (-9.56 kJ/mol),[285] which is in fairly good agreement with the "POL+CT" value given by ALMO-EDA. Moreover, it was suggested by Řezáč et al. that the δHF term contains a significant contribution from CT as it manifests a good correlation with CT energies computed with CDFT(FBH), [341] which further supports our perspective that the relatively small magnitude of regularized SAPT CT is at least partly caused by not taking the δHF term (high-order induction effect) into account in the formulation of this theory.

4.4 Conclusion

In this paper, we have conducted a thorough study on the definitions of CT in the context of intermolecular interactions. The performance of the ALMO- and CDFT-based approaches, which represent definitions based on orbital- and real-space partitioning, respectively, is investigated and compared on a wide range of model systems concerning the energetic stabilization effect (ΔE_{CT}) as well as the observable consequences. Techniques recently developed by us, such as the constant-density SCF-MI/SCF calculations and the adiabatic EDA scheme (the latter is extended to CDFT in this work), are utilized in this study to provide useful information. The main findings are summarized as follows:

- The CDFT definition concerns the net charge flow between molecules, while the ALMO and many other orbital-based definitions regard CT as an intermolecular orbital relaxation

effect. Therefore, the "CT-free" state obtained via CDFT can be contaminated by energy lowering due to intermolecular relaxation as long as forward and backward donations shift equal fractions of an electron. This effect is most pronounced for the cyclic H-bonding systems where the transfers in two ways exactly cancel each other because of symmetry, and it is also revealed in other systems (BH_3-CO, Ni(CO)) that involve a significant back-donation.

- Significant intermolecular relaxation can take place without changing the fragment populations, i.e., the variational principle is able to work around the constraints in a CDFT calculation. This is revealed by calculating the constant-density CT energy, which gives a variational upper bound to the energy lowering (negative) associated with the population-conserving relaxation. Such an effect explains why CDFT usually gives smaller CT energies than ALMO's even for systems whose CT is presumably unidirectional, as illustrated by examples such as the water dimer and the NH_3-BH_3 complex.

- The above two differences extend to the observable consequences of CT calculated with the ALMO and CDFT definitions. It is a probable signature of CT-contamination that many well-established fingerprints of CT, such as the red shift in water's OH stretch frequency when it is bound to Cl^-, and the bending of the planar BH_3 upon the formation of donor-acceptor complexes, already appear in the "CT-free" state constructed by CDFT. On the other hand, the changes in these observables are characterized as effects of CT through adiabatic EDA calculations using the ALMO definition.

- There exist special cases where CDFT yields larger CT energies than the ALMO results. In the cases of water-Mg^{2+}, Li^+, Na^+ complexes examined above, the CDFT results showed a strong dependence on the choice of population scheme. Evidence was provided that the CDFT(Becke) scheme, which yields anomalously large values of CT, are likely to be inappropriate for these cases.

It should be acknowledged that the CDFT-based definition does have a number of desirable features. For instance, the CT energies calculated by CDFT are fairly insensitive to the employed basis set, which is an advantage over earlier orbital-based approaches. From the practical perspective, the CDFT definition of CT might be favored for the development of molecular force fields, as it is clearly difficult to use an empirical function to model a term of large magnitude concerning quantum delocalization effect. CT energies computed by CDFT, in principle, should be better correlated with net population change upon complexation, where the latter can be measured for certain systems by experimental means.[361]

The results of this paper highlight the fact that the ALMO and CDFT definitions of CT capture different physics. The CDFT condition prohibits net population transfer between fragments, and thus its constrained "CT-free" state is demonstrably CT-contaminated in any system where CT can occur without net population change. The ALMO-based definition aims to prohibit prohibit all intermolecular orbital mixing, and therefore is more consistent

with the common understanding of CT effects in chemistry, especially in terms of the calculated observable consequences of CT. The recently proposed FERF approach renders the basis set limit of the resulting CT energy well-defined while retaining the preferable features of the ALMO definition. Therefore, it can be recommended for characterizing the effects of CT in non-covalently bound systems. Development of the associated theoretical tools, such as a pairwise CT analysis scheme (which is analogous to the original ALMO-CTA [305]) and the analytical nuclear gradient for the "SCF-MI(FERF)" PES (for performing adiabatic EDA), will be highly desirable, and progress has been made in these directions.

Chapter 5

Assessment of the AMOEBA Force Field

5.1 Introduction

Condensed-phase simulations with molecular dynamics or Monte Carlo methods afford the ability to probe physical properties not easily accessible by experimental means, but only when there is a reliable model of the potential energy surface. However for simulations on scales representing tens of thousands to millions of atoms, quantum mechanical (QM) methods become extremely intractable, if not impossible, to serve that purpose. Therefore, there exists a keen interest in reducing the computational cost by representing such systems classically, and considering moreover only the nuclear degrees of freedom explicitly under the Born-Oppenheimer separation of the timescales of the electronic and nuclear degrees of freedom.

In molecular mechanics (MM) formulations, a potential energy function of the positions of the atomic nuclear positions is used to describe covalent interactions (bond stretching, angle bending, and torsional rotation, and sometimes other cross-terms) as well as non-covalent interactions such as van der Waals (vdW) and point-charge permanent electrostatics. Fundamentally the MM description involves idealized functional forms for these terms such as

$$U = \sum_{b}^{\text{bonds}} k_b(b-b_0)^2 + \sum_{\theta}^{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\phi}^{\text{torsions}} k_\phi[1+\cos(n\phi+\delta)] + \sum_{i<j} 4\epsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \sum_{i<j}\frac{q_i q_j}{r_{ij}} \tag{5.1}$$

For bonded atoms, Eq. (5.1) permits only small fluctuations around the equilibrium bond-lengths and bond angles. It employs nuclear-centered point charges for the permanent electrostatics, and uses simple pairwise additivity and mixing rules for the vdW interactions. It has also become more common to replace point charges with an extended set of point

multipoles, and to include non-pairwise additive interactions such as polarization which is discussed in detail below.

While the long-range asymptotic behavior of these functional forms are correctly modeled for electrostatics and dispersion[94], the short-ranged asymptotics in the compressed region are more approximate, for example in the steep region of the van der Waals function which assumes a power law for the Pauli exclusion term as opposed to the exponential form from QM. Other short-ranged QM features such as a non-local electron density extent from the nucleus, which is related to the charge penetration (CP) effect, or interactions such as charge transfer (CT), are ignored altogether in Eq. (5.1), although very recently approximate classical formulations to these terms are just starting to appear in the literature.[142, 153, 362–376]

Although the classical functional forms for the Born-Oppenheimer potential energy surface (BO-PES) are largely correct, and are starting to approach a level of completeness for all relevant QM interactions, the methods by which the parameters are obtained for these functions can be more arbitrary. This is especially true for the soft, non-covalent degrees of freedom that fluctuate much more in distance than that experienced due to the covalent interactions at typical ambient conditions. For example, deriving point charge or point multipole parameters by fitting to the QM-derived electrostatic potential at a series of distances outside of the vdW surface is often a poorly defined problem, with charges or multipoles of buried atoms far from the surface being able to vary unphysically and still reproduce the potential[377], and such problems are compounded when including the additional polarization terms. The vdW parameters are especially difficult to fit since the attractive part of the vdW interaction, the dispersion interaction, cannot be fitted easily to QM data owing to the difficulty of finding accurate yet tractable levels of theory that adequately capture electron correlation[92, 378, 379]. Together with the repulsive wall, the overall vdW function is typically tuned through a reliance on experimental data, by adjusting the parameters to reproduce densities and heats of vaporization of neat liquids. [139, 140, 380–382] In typical force field development, parameters for the individual terms in Eq. (5.1) are determined independently and then refined together in an iterative fashion to capture missing features like charge penetration, polarization if it is absent, or even directly targeting better properties for a particular phase (such as liquid water)[383].

Electronic structure calculations have been routinely used to benchmark the quality of force field parameterization, since they facilitate the direct comparison of resulting interaction energies and other properties within a given model system. Nevertheless, considering the piecemeal parameterization of a force field based on several minimally correlated approximating assumptions and the prevalence of incomplete error cancellation between terms in the final form, it would be helpful to benchmark against a method that is able to ascertain the quality of individual terms of the force field. Energy decomposition analysis (EDA) affords a way to determine the relative contributions of several physically meaningful terms out of the QM interaction energy, e.g., permanent electrostatics, Pauli repulsion, polarization, dispersion, etc.[98, 105, 109] The asymptotic components of an EDA method are uniquely defined for a given electronic structure method. Furthermore, despite some arbitrariness

that exists in the definitions of these terms (since none of them are true experimental observables), well-defined EDAs can yield a reasonable and chemically sensible separation of energy components in the overlapping regime. Therefore, by comparing the corresponding terms between an EDA scheme and a force field, one can in principle obtain insight into the strengths and weaknesses of MM formulations.

In this work, we analyze the energy decomposition of the popular polarizable MM force field AMOEBA. [139–141, 384] The AMOEBA model goes beyond typical fixed-charge force field by including both higher-order permanent multipoles and inducible dipoles, in an attempt to more faithfully reproduce the BO-PES from more accurate QM methods. As such, AMOEBA results can be directly compared to and parameterized against a corresponding BO-PES derived from *ab initio* electronic structure calculations, while remaining compatible with the goal of increased computational efficiency and ease of parameterization for a variety of systems including ions[310, 385], small organic molecules[140], transition metal complexes [386, 387], proteins[388], and nucleic acids either in the gas or condensed phases. The subject of this paper is to assess how well the individual energy terms of the AMOEBA model are recapitulated when compared to QM on the simplest of systems — namely the water dimer and a series of water-ion dimers — over a range of distances and angular orientations.

Of the available energy decomposition schemes applied to force field development, two of the more well known examples include symmetry-adapted perturbation theory (SAPT) [93, 142, 366, 389–395] and variational based EDA approaches [97, 115, 308, 310, 328, 389, 396, 397]. SAPT evaluates intermolecular interactions via a perturbative approach[92, 334, 379], and the resulting interaction energies are decomposed into the contributions of electrostatics, exchange, induction, and dispersion. The development of SAPT(DFT) [398–401] or DFT-SAPT [402–404], which tackles intramolecular correlation by means of Kohn-Sham (KS)-DFT [5], offers a reasonable balance between accuracy and efficiency ($O(N^5)$ scaling with the use of density-fitting[405, 406]) so that it can be used for generating the training data for force field parametrization.[93] While SAPT-based methods facilitate the development of many advanced force fields (the readers are referred to the review in Ref. 93), it still faces several challenges. Most importantly, the accuracy of the total interaction energy relies on the convergence of the perturbative expansion, which is often difficult to assure when methods like DFT-SAPT are used for systems with strong induction effects.[75, 100] Also, there is no clean separation between polarization and charge transfer in the conventional formulation of SAPT (they both belong to the induction term), although several approaches have been proposed to extract the CT energy.[101–103]

We shall evaluate AMOEBA using the variational absolutely localized molecular orbital (ALMO)-EDA scheme[97, 115, 116], which partitions the total intermolecular interaction energy into contributions of frozen orbital interaction (which contains permanent electrostatics, Pauli repulsion, and dispersion), polarization and CT. New advances made in the ALMO-EDA scheme include (1) the ability to reach a meaningful complete basis set (CBS) limit for polarization and CT using the fragment electric-field response function (FERF) model[117] and (2) the ability to disentangle the contributions from the aforementioned three components of the frozen term[118], which is not further separated in the original scheme.

These advances define a second generation ALMO-EDA[97] method which is employed in the present paper.

Apart from the robustness of this EDA scheme, another important motivation for employing a DFT-based EDA approach is the significantly reduced errors of state-of-the-art density functionals for non-covalent interactions.[27, 28, 78, 407, 408] It is noteworthy that the functionals recently developed by Mardirossian and Head-Gordon[72–74] demonstrate very good accuracy for noncovalent interactions when large basis sets are employed. Based on recent benchmark by Lao et al.[75], the mean absolute error (MAE) of ωB97X-V [72] (using the aug-cc-pVTZ (aTZ) [178, 179] basis set) for a composite dataset comprising neutral-anion, neutral-cation, and cation-anion interactions (43 data points in total) is 0.55 kcal/mol, which is superior to that of the popular DFT-SAPT method (1.43 kcal/mol) and comparable to the most accurate SAPT result available (0.43 kcal/mol, as computed at the SAPT2+3-δMP2/aTZ level of theory). We note that such high-level SAPT methods are computationally costly ($O(N^7)$ scaling) and offer a less unambiguous energy decomposition due to the coupling between terms. Beyond equilibrium binding energies, accurate PESs generated by ωB97X-V for water-water [78] and water-anion (F^-, Cl^-) dimers [76] have also been recently reported. Therefore, we use ωB97X-V for all the ALMO-EDA calculations in the present paper.

The remainder of this paper is outlined as follows. Pertinent details of the AMOEBA force field are summarized in Sec. 5.2.1, followed in Sec. 5.2.2 by a concise summary of the version of the ALMO-EDA that will be applied in this work and the mapping between its terms and those in AMOEBA. The resulting data and analysis for four categories of interactions are demonstrated and discussed in Sec. 5.3. For each category, we first compare total energies evaluated with ωB97X-V and AMOEBA, and assess the agreement near the equilibrium configuration, as well as at short-range (the so-called compressed region) and long-range (the asymptotic region). We then compare the relative contributions of the energetic components, including permanent electrostatics, polarization, and vdW interactions, in various regions, via which we elucidate why the resultant total interaction energy profile given by AMOEBA for each system has satisfactory or poor agreement with the ωB97X-V result. Moreover, in cases where total interaction energies are in reasonable agreement between DFT and AMOEBA, we investigate how the effects of CP and CT might be accounted for implicitly by AMOEBA, as it lacks explicit functional forms for these effects. The insights gained from these benchmark calculations are discussed in Sec. 5.4.

5.2 Computational Methods

5.2.1 Non-covalent terms in the AMOEBA force field

The non-covalent (intermolecular) terms in the AMOEBA force field comprises permanent electrostatics ($U_{\text{ele}}^{\text{perm}}$), induced electrostatics ($U_{\text{ele}}^{\text{ind}}$), and van der Waals (vdW) interactions (U_{vdw}). An atom-centered point multipole model is adopted for permanent electrostat-

ics: on each atomic site i, the (vector) of permanent multipoles $\mathbf{M}_i$ includes monopole (q_i), dipole ($\boldsymbol{\mu}_i$), and quadrupole ($\mathbf{Q}_i$) moments:

$$\mathbf{M}_i^T = [q_i, \mu_{ix}, \mu_{iy}, \mu_{iz}, Q_{ixx}, Q_{ixy}, Q_{ixz}, Q_{iyy}, Q_{iyz}] \tag{5.2}$$

The total permanent electrostatics contribution is then evaluated as the pairwise sum of interactions between different atomic sites:

$$U_{\text{ele}}^{\text{perm}} = \sum_{i<j} \mathbf{M}_i^T \mathbf{T}_{ij} \mathbf{M}_j \tag{5.3}$$

where $\mathbf{T}_{ij}$ is the "composite" multipole interaction tensor between sites i and j, that contain appropriate powers of $1/r_{ij}$ according to the permanent multipole expansion for the AMOEBA potential[139, 384]. For water and simple ions, the interaction sites i and j are constrained to be on different molecular or atomic fragments; and for the mono-atomic ions studied in this work, the RHS of Eq. (5.2) contains monopole (q_i) only. The set of multipoles are derived from a QM electronic density using the distributed multipole analysis (DMA)[409], whose values are further refined by a fit to an electrostatic potential generated by a higher level of QM theory [140].

The polarization effect in AMOEBA is modeled by induced dipoles, $\boldsymbol{\mu}_i^{\text{ind}}$ placed on each atomic site i, whose magnitude is determined by the site-specific isotropic polarizability and the total external electric field exerted:

$$\boldsymbol{\mu}_i^{\text{ind}} = \alpha_i (\mathbf{E}_i + \mathbf{E}_i') \tag{5.4}$$

where $\mathbf{E}_i$ is the electric field owing to the permanent multipoles on other fragments, and $\mathbf{E}_i'$ is the field generated by the induced dipoles on all the other atomic sites:

$$\mathbf{E}_i = \sum_j \mathbf{T}_{ij} \mathbf{M}_j^{(d)} \tag{5.5}$$

$$\mathbf{E}_i' = \sum_{j \neq i} \mathbf{T}_{ij}' \boldsymbol{\mu}_j^{ind} \tag{5.6}$$

where $\mathbf{T}_{ij}'$ now refers to appropriate powers of $1/r_{ij}$ according to the dipole induction and the superscript (d) refers to special scaling factors used for electrostatic interactions in AMOEBA[140] (not involved in this work). Since the RHS of Eq. (5.4) relies on the induced dipoles, $\{\boldsymbol{\mu}_i^{\text{ind}}\}$ are solved self-consistently in order to complete the calculation of the many-body polarization effects. With converged $\{\boldsymbol{\mu}_i^{\text{ind}}\}$, the associated energy lowering (polarization energy) is determined by

$$U_{\text{ele}}^{\text{ind}} = -\frac{1}{2} \sum_i \boldsymbol{\mu}_i^{\text{ind}} \cdot \mathbf{E}_i \tag{5.7}$$

The atomic polarizability parameters, $\{\alpha_i\}$, are derived by a fit to available experimental molecular polarizabilities[139, 384].

One artifact of the distributed interactive induced electrostatics model is the so-called "polarization catastrophe", i.e., the electric field generated by point multipoles can severely overpolarize at short range and even lead to divergence. To ensure the finite nature of the intermolecular induction effect, a Thole-style damping scheme is employed by AMOEBA [410, 411], which is equivalent to replacing a point multipole with a smeared charge distribution. For example, the damping function for monopoles has the following form:

$$\rho = \frac{3a}{4\pi}\exp(-au^3), \quad u = \frac{r_{ij}}{(\alpha_i\alpha_j)^{1/6}} \tag{5.8}$$

where r_{ij} is the distance between sites i and j, α_i, α_j are their polarizabilities, and a is a dimensionless width parameter. The damping functional forms for higher multipoles are reported in Ref. 139. In practice, the damping functions are built in the formation of multipole interaction tensors in Eq. (5.5) and (5.6).

For the vdW interaction, AMOEBA adopts a pairwise additive buffered 14-7 potential originally proposed by Halgren [412]:

$$U_{\text{vdw}} = \sum_{i<j}\epsilon_{ij}\left(\frac{1+\delta}{\rho_{ij}+\delta}\right)^7\left(\frac{1+\gamma}{\rho_{ij}^7+\gamma}-2\right) \tag{5.9}$$

where ϵ_{ij} is the depth of the potential well, ρ_{ij} is the dimensionless distance between sites i and j: $\rho_{ij} = r_{ij}/R^0_{ij}$, where R^0_{ij} is the minimum energy separation, and γ and δ are two constants whose values are set to 0.12 and 0.07, respectively. If we further expand the RHS of Eq. (5.9), the repulsive "14" term ($U^{\text{rep}}_{\text{vdw}}$) mostly accounts for Pauli repulsion, while the attractive "7" term ($U^{\text{attr}}_{\text{vdw}}$) in principle yields a more accurate series expansion for dispersion. In the current parameterization regime of AMOEBA, the homonuclear vdW parameters ϵ_{ii} and R^0_{ii} are obtained by a fit to dimer energies calculated at the MP2/aTZ level of theory or higher, followed by a refinement stage where experimental liquid densities and heats of vaporization are reproduced [140]; and the heteronuclear parameters (ϵ_{ij} and R^0_{ij}) are obtained by using the combination rules. For more details we refer the reader to the original AMOEBA references [139–141, 384].

5.2.2 Energy decomposition analysis

The initial partitioning of the total interaction energy in the ALMO-EDA method[97, 115, 116] is as follows:

$$E_{\text{int}} = E_{\text{gd}} + E_{\text{frz}} + E_{\text{pol}} + E_{\text{ct}}. \tag{5.10}$$

The first term (geometry distortion) describes the energy change due to the geometric distortion of monomers to the complex structure from their fully relaxed geometries, which is *not* considered in this work (in AMOEBA it is captured by the bonded terms). The frozen orbital interaction, E_{frz}, is defined as the energy difference between the frozen orbital wavefunction (which corresponds to a frozen one-particle density matrix $\mathbf{P}_{\text{frz}}$) and the sum of

monomer energies that are computed individually:

$$E_{\text{frz}} = E(\mathbf{P}_{\text{frz}}) - \sum_A E_A(\mathbf{P}_A), \tag{5.11}$$

It represents the energy change when fragments approach each other without any variational relaxation of their orbitals or density, apart from ensuring that they obey the Pauli Principle.

The frozen interaction can be further decomposed into contributions from permanent electrostatics, Pauli repulsion and dispersion interactions. Our original approach [118] is based on the partitioning of $\mathbf{P}_{\text{frz}}$ into a sum of fragment contributions: $\mathbf{P}_{\text{frz}} = \sum_A \tilde{\mathbf{P}}_A$, where interfragment orthogonality is enforced between $\tilde{\mathbf{P}}_A$'s. This is achieved via a constrained minimization of the "kinetic energy pressure" (KEP) objective function, as described in Ref. 118.

$\tilde{\mathbf{P}}_A$ can be regarded as the deformed (but not yet relaxed) density of each individual fragment upon the formation of the initial complex. The dispersion energy is then defined as the remainder of interfragment exchange-correlation (XC) energy after subtracting the part that is "dispersion-free" (DF) in nature:

$$\begin{aligned} E_{\text{disp}} = &\left(E_{\text{xc}}[\mathbf{P}_{\text{frz}}] - \sum_A E_{\text{xc}}[\tilde{\mathbf{P}}_A] \right) \\ &- \left(E_{\text{xc}}^{\text{DF}}[\mathbf{P}_{\text{frz}}] - \sum_A E_{\text{xc}}^{\text{DF}}[\tilde{\mathbf{P}}_A] \right), \end{aligned} \tag{5.12}$$

For this purpose, an auxiliary DFXC functional is required, and our previous work [97, 118] suggests that Hartree-Fock (HF) theory is an appropriate choice for dispersion-corrected range-separated hybrid (RSH) functionals like ωB97X-V.

The approach presented in Ref. 118 utilizes $\{\tilde{\mathbf{P}}_A\}$ to define permanent electrostatics and Pauli repulsion as well, i.e., all three energy components are computed making use of the properly antisymmetrized wavefunction. That approach properly describes the deformation of monomer densities due to Pauli repulsion (antisymmetrization of the supersystem wavefunction) [118]. However such deformations cannot be captured by AMOEBA or any other force field whose permanent electrostatics is described by multipole moments that are invariant with intermolecular separations. Therefore, in order to make ALMO-EDA's permanent electrostatics physically compatible with AMOEBA, we *step back* to adopt the "classical electrostatics" definition[104–107], which describes the Coulomb interaction between charge distributions of isolated fragments:

$$E_{\text{elec}}^{\text{cls}} = \sum_{A<B} \int_{r_1} \int_{r_2} \rho_A^{\text{tot}}(\mathbf{r}_1) \frac{1}{r_{12}} \rho_B^{\text{tot}}(\mathbf{r}_2) \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \tag{5.13}$$

where $\rho_A^{\text{tot}}(\mathbf{r}) = \rho_A^{\text{el}}(\mathbf{r}) + \rho_A^{\text{nuc}}(\mathbf{r})$. The modified Pauli term is then simply defined as the remainder of the frozen interaction, which still comprises the repulsive interaction stemming from Pauli exclusion principle and interfragment exchange effects. Taken together, the

decomposition of the frozen term adopted in this work can be expressed as

$$E_{\text{frz}} = E_{\text{elec}}^{\text{cls}} + E_{\text{Pauli}}^{\text{mod}} + E_{\text{disp}} \tag{5.14}$$

The contribution of polarization is determined by variationally optimizing the density matrix associated with a fragment-block-diagonal (absolutely localized) MO coefficient matrix, using locally projected self-consistent field (SCF) techniques [233, 234, 301] or a gradient-based minimization. [117] This procedure is called "SCF for molecular interactions" (SCF-MI). The resulting ALMO density matrix, $\mathbf{P}_{ALMO}$, is employed to evaluate the polarization energy:

$$E_{\text{pol}} = E[\mathbf{P}_{\text{ALMO}}] - E[\mathbf{P}_{\text{frz}}] \tag{5.15}$$

The use of ALMOs ensures that the net charge on each fragment is conserved under the Mulliken population definition, i.e., CT between fragments is not allowed. The contribution of CT is then defined as the energy difference between this "CT-forbidden" SCF-MI wavefunction and the fully relaxed one:

$$E_{\text{ct}} = E[\mathbf{P}_{\text{SCF}}] - E[\mathbf{P}_{\text{ALMO}}] \tag{5.16}$$

In practice, the AO-based fragment partitioning used in SCF-MI breaks down when very large AO basis sets (especially those with diffuse functions) are used, which results in an overestimated polarization energy (contaminated by CT). [103, 117, 316] In order to judiciously choose the degrees of freedom that are relevant to polarization, Horn and Head-Gordon introduced the fragment electric-field response function (FERF) model[117] which defines the fragment subspaces based on the response of MOs to an external electric field (and its spatial derivatives). The FERFs are able to capture the relaxation of fragment occupied orbitals under a weak external field, which is deemed as the physical essence of polarization (see Ref. 117 for more details on how FERFs are constructed). In this work, we use the "FERF-nDQ" model (non-orthogonal FERFs accounting for dipole and quadrupole responses, which require 3 and 5 FERFs for each occupied orbital, respectively) to compute the polarization energy instead of the original AO-based approach. The FERF-nDQ model appears to give a satisfactory description[97, 117] of polarization effects with a well-behaved basis set limit. The equations utilized to determine the contributions of polarization and CT are identical to Eqs. (5.15) and (5.16) in form.

Here we also briefly compare the intermolecular interaction components generated by ALMO-EDA with those from a standard DFT-SAPT calculation, due to the popularity of the latter approach in developing physically-motivated force fields. The DFT-SAPT interaction energy can be written as (following the notation of Ref. 100):

$$\begin{aligned} E_{int}^{\text{DFT-SAPT}} &= [E_{\text{elst}}^{(1)}]_{\text{elst}} + [E_{\text{exch}}^{(1)}]_{\text{exch}} \\ &+ [E_{\text{ind}}^{(2)} + E_{\text{exch-ind}}^{(2)} + \delta E_{\text{HF}}^{(2)}]_{\text{ind}} \\ &+ [E_{\text{disp}}^{(2)} + E_{\text{exch-disp}}^{(2)}]_{\text{disp}} \end{aligned} \tag{5.17}$$

Subscripts outside the brackets indicate the grouping of terms into four energy components: electrostatics, exchange, induction and dispersion. While the first-order electrostatic term is identical to ALMO-EDA's (classical) electrostatics (Eq. (5.13)) in form, the other components are computed differently in DFT-SAPT due to its perturbative approach. Nonetheless, based on the physical meaning of each term, there exists a clear correspondence between the terms in DFT-SAPT (left) and ALMO-EDA (right):

- Exchange → Pauli repulsion
- Induction → Polarization + CT
- Dispersion → Dispersion

Numerically their resulting energy components should be comparable at least to some extent. As an example, we compare the results of DFT-SAPT (provided in Ref. 285) and ALMO-EDA for CCSD(T)[96]-optimized structures of the linear and bifurcated water dimer (taken from Ref. 306), which are presented in Table D.1 (in Appendix D). In general, there is *no* qualitative difference between the results of these two decomposition schemes (in contrast, the difference between the energy components of AMOEBA and either EDA scheme can be much larger), while the separation of polarization and CT in ALMO-EDA seems to further facilitate the comparison with AMOEBA.

Finally, the correspondence between terms in AMOEBA and the ALMO-EDA scheme used in this work is summarized in Table. 5.1. Note that in the following discussion, we use "total electrostatics" to represent the sum of permanent electrostatic interactions and polarization (induced electrostatics); and "vdW interaction" refers to the entire 14-7 potential for AMOEBA, while for ALMO-EDA it refers to the sum of modified Pauli repulsion and dispersion.

Table 5.1: Correspondence of terms in AMOEBA and ALMO-EDA

Components	AMOEBA	ALMO-EDA
Permanent Electrostatics	$U_{\text{ele}}^{\text{perm}}$	$E_{\text{elec}}^{\text{cls}}$
Induced Electrostatics	$U_{\text{ele}}^{\text{ind}}$	E_{pol}
Pauli Repulsion	$U_{\text{vdw}}^{\text{rep}}$	$E_{\text{Pauli}}^{\text{mod}}$
Dispersion	$U_{\text{vdw}}^{\text{attr}}$	E_{disp}
Charge Transfer	no explicit	E_{ct}

5.2.3 Computational details

Energy calculations using the AMOEBA force field were performed in the Tinker7 molecular modeling package [413]. The most recently released parameters were used for all species: "amoebapro13" was used for water-water, water-cation, and water-Cl^- interactions, while

for water-F^- and water-Br^-, parameters were obtained from "amoeba09", which corresponds to the latest parameterization for these two halides [139, 140, 309, 385, 388, 414]. Neither periodic boundary conditions nor distance cutoffs were adopted for any of these calculations; therefore, permanent electrostatics and polarization were performed in a standard, no-cutoff direct space interaction. Induced dipoles were converged to 10^{-12} Debye. For the calculation of the buffered 14-7 vdW potential, a trivial modification to the source code was made to allow the repulsive and attractive terms to be reported separately.

All the ALMO-EDA calculations in this work were performed with a standard version of the Q-Chem 4.4 software package [255]. The ωB97X-V functional,[72] which is a range-separated hybrid GGA that incorporates the VV10 [71] non-local correlation functional for the description of dispersion, is used for modeling the intermolecular interactions. The large def2-QZVPPD[183] basis set (augmented quadruple-ζ) is employed *without* counterpoise corrections [88] for basis set superposition errors (BSSE), since for dimer interactions, the BSSE associated with the use of this basis set should be almost negligible compared to the magnitude of the investigated interactions [87]. The numerical integration of the XC functional is performed on a $(99, 590)$ grid (99 radial shells with 590 Lebedev points on each), while the SG-1 grid [258] is used to integrate the non-local correlation functional.

The frozen energy decomposition is based on the modified scheme defined by Eq. (5.14), using Hartree-Fock as the DF functional for the separation of dispersion. The polarization contribution is determined through the "FERF-nDQ" model, as introduced in Sec. 5.2.2. All the variational energy minimizations involved in ALMO-EDA (SCF on isolated fragments, SCF-MI, and supersystem SCF) are converged to 10^{-8} a.u..

For the PES scans, we start from the equilibrium geometry optimized at the ωB97X-V/def2-QZVPPD level of theory. Then, we stretch/compress the complex along one single chosen coordinate, *without* relaxing other degrees of freedom (a rigid dissociation/compression). Unless otherwise specified, for all the distance scans, the separation between the oxygen in water and the ion ($O\cdots O$ distance for the water dimer case) is selected as the coordinate being modified, and the interval between neighboring data points is 0.05 Å. AMOEBA and ALMO-EDA results are then generated on the same set of configurations.

In order to further validate the QM model chemistry employed in this work, we compare the interaction energies of the water dimer and five water-ion dimers (at compressed, equilibrium and stretched geometries) evaluated by ωB97X-V/def2-QZVPPD and AMOEBA against the ΔCCSD(T)/CBS reference values (Table 5.2). More computational details for this benchmark are provided in the caption of Table 5.2. ωB97X-V provides sub-kcal/mol accuracy for almost all the investigated systems except for the stretched configuration of $H_2O\cdots Mg^{2+}$. ωB97X-V tends to slightly overbind the water-cation complexes in the long range due to self-interaction error, which is most pronounced in $H_2O\cdots Mg^{2+}$. The development of ALMO-EDA for wavefunction-based correlation methods [120, 284] could provide useful alternatives in such scenarios. The error of ωB97X-V is typically over 5 times smaller than AMOEBA's so that we can use ALMO-EDA to trace the source of error among AMOEBA's energy components. In cases where the ratio in errors is smaller (e.g., for the

Table 5.2: Total interaction energies (in kJ/mol) of the water dimer and five water-ion dimers evaluated at equilibrium, compressed (10%) and stretched (10%) configurations, where the compressions and stretches are applied to the same coordinates as in the rigid PES scans. For the ΔCCSD(T)/CBS reference, core-valence correlated Dunning basis sets aug-cc-pCVTZ and aug-cc-pCVQZ [415, 416] are employed for a two-point extrapolation of the MP2 correlation energies [417], and the correction for higher-order correlation (E[CCSD(T)] - E[MP2]) is computed at the aug-cc-pCVTZ level. All the correlation energies are computed without the frozen-core approximation. As opposed to the DFT calculations performed in this work, counterpoise corrections for BSSE are applied when computing the ΔCCSD(T)/CBS interaction energies.

		ωB97X-V	AMOEBA	Ref.	Error(ωB97X-V)	Error(AMOEBA)
Compressed	$H_2O\cdots H_2O$	-14.31	-13.70	-14.86	0.55	1.16
	$H_2O\cdots Li^+$	-133.37	-127.34	-133.72	0.35	6.38
	$H_2O\cdots Na^+$	-89.49	-84.97	-89.81	0.32	4.84
	$H_2O\cdots Mg^{2+}$	-317.40	-298.55	-317.38	-0.02	18.83
	$H_2O\cdots F^-$	-106.43	99.68	-108.35	1.92	208.03
	$H_2O\cdots Cl^-$	-52.32	-43.93	-52.77	0.45	8.84
Equilibrium	$H_2O\cdots H_2O$	-21.07	-21.35	-21.25	0.18	-0.10
	$H_2O\cdots Li^+$	-146.27	-139.26	-145.25	-1.02	5.99
	$H_2O\cdots Na^+$	-101.69	-97.83	-100.70	-0.99	2.87
	$H_2O\cdots Mg^{2+}$	-347.79	-326.19	-344.86	-2.93	18.67
	$H_2O\cdots F^-$	-132.79	-97.18	-133.12	0.33	35.94
	$H_2O\cdots Cl^-$	-64.79	-68.91	-64.72	-0.07	-4.19
Stretched	$H_2O\cdots H_2O$	-18.53	-17.86	-18.49	-0.04	0.63
	$H_2O\cdots Li^+$	-139.40	-131.87	-137.69	-1.71	5.82
	$H_2O\cdots Na^+$	-95.85	-92.53	-94.36	-1.49	1.83
	$H_2O\cdots Mg^{2+}$	-331.64	-302.60	-326.85	-4.79	24.25
	$H_2O\cdots F^-$	-122.08	-128.31	-121.87	-0.21	-6.44
	$H_2O\cdots Cl^-$	-59.55	-63.05	-59.19	-0.36	-3.86

equilibrium water dimer, the error of AMOEBA is even smaller than that of ωB97X-V), the interaction energy given by AMOEBA is usually fairly accurate, and ALMO-EDA can be utilized to uncover the origins of such well-behaved cases.

5.3 Results

The equilibrium intermolecular distances and interaction energies for all the studied dimer complexes are summarized in Table 5.3. Note that the AMOEBA interaction energies in Table 5.3 are evaluated at MM-relaxed geometries so they do *not* correspond to any points on the potential energy curves in the figures presented below. We also note that the intermolecular interaction energies of $H_2O\cdots H_2O$ and $H_2O\cdots Cl^-$ given by AMOEBA turn out to be *less favorable* at their AMOEBA-optimized geometries than at the QM minima (the interaction energies for the latter are given in Table 5.2), which, nevertheless, are compensated by the intramolecular relaxation of the involved water monomers.

Table 5.3: Equilibrium intermolecular separations (Å) and total interaction energies (kJ/mol) for the water dimer and eight water-ion dimers, based on geometries optimized at the ωB97X-V/def2-QZVPPD level of theory (left) and with the AMOEBA force field (right). The distance from the oxygen atom of the water molecule to the ion in each complex (O···O distance for the water dimer case) is reported.

	ωB97X-V		AMOEBA	
	distance	E_{int}	distance	E_{int}
$H_2O\cdots H_2O$	2.92	-21.07	2.89	-20.86
$H_2O\cdots Li^+$	1.84	-146.27	1.82	-140.67
$H_2O\cdots Na^+$	2.22	-101.69	2.23	-98.04
$H_2O\cdots K^+$	2.62	-74.10	2.60	-73.16
$H_2O\cdots Mg^{2+}$	1.91	-347.79	1.88	-332.87
$H_2O\cdots Ca^{2+}$	2.22	-242.55	2.22	-228.65
$H_2O\cdots F^-$	2.45	-132.79	2.64	-119.81
$H_2O\cdots Cl^-$	3.12	-64.79	3.15	-66.44
$H_2O\cdots Br^-$	3.31	-55.95	3.36	-55.07

5.3.1 The water dimer

We first assess the performance of AMOEBA for the water dimer interaction against the QM results. Since AMOEBA was initially designed as an advanced polarizable water model, we expect it to give a high-quality depiction of the PES for the water dimer, the prototypical system for water-water interactions. The top left panel of Figure 5.1 shows an overall good match between the potential energy curve evaluated by AMOEBA and ωB97X-V. For a rigid dissociation curve, AMOEBA and ωB97X-V predict the same O···O distance for the energy minimum at 2.90 Å, with a minimal discrepancy in energy: AMOEBA is slightly more bound by -0.35 kJ/mol, a difference that is close to the intrinsic error of the functional for this system. The agreement between the two potential energy curves near the equilibrium separation is more clearly demonstrated by the inset plot. In the highly compressed region (< 2.70 Å), AMOEBA gives a distinctly harder repulsive wall. Its long-range interaction is also less attractive than the ωB97X-V profile, although the discrepancy is fairly small (the maximum deviation is about 1 kJ/mol around $R_{\text{O}\cdots\text{O}} = 3.6$ Å).

The component breakdowns of QM and AMOEBA interaction energies in the same range are plotted in the middle and bottom left panels. The permanent electrostatic component of AMOEBA is less attractive compared to its "classical electrostatics" counterpart in ALMO-EDA at short range, although they agree in the asymptotic region. By contrast close agreement for the polarization energy is achieved over all intermolecular separations except for the highly compressed region ($R_{\text{O}\cdots\text{O}} < 2.6$ Å), where the polarization energy of AMOEBA becomes less favorable than that given by ALMO-EDA due to the onset of Thole damping in the former. The profiles for total electrostatics mostly reflect the above mentioned

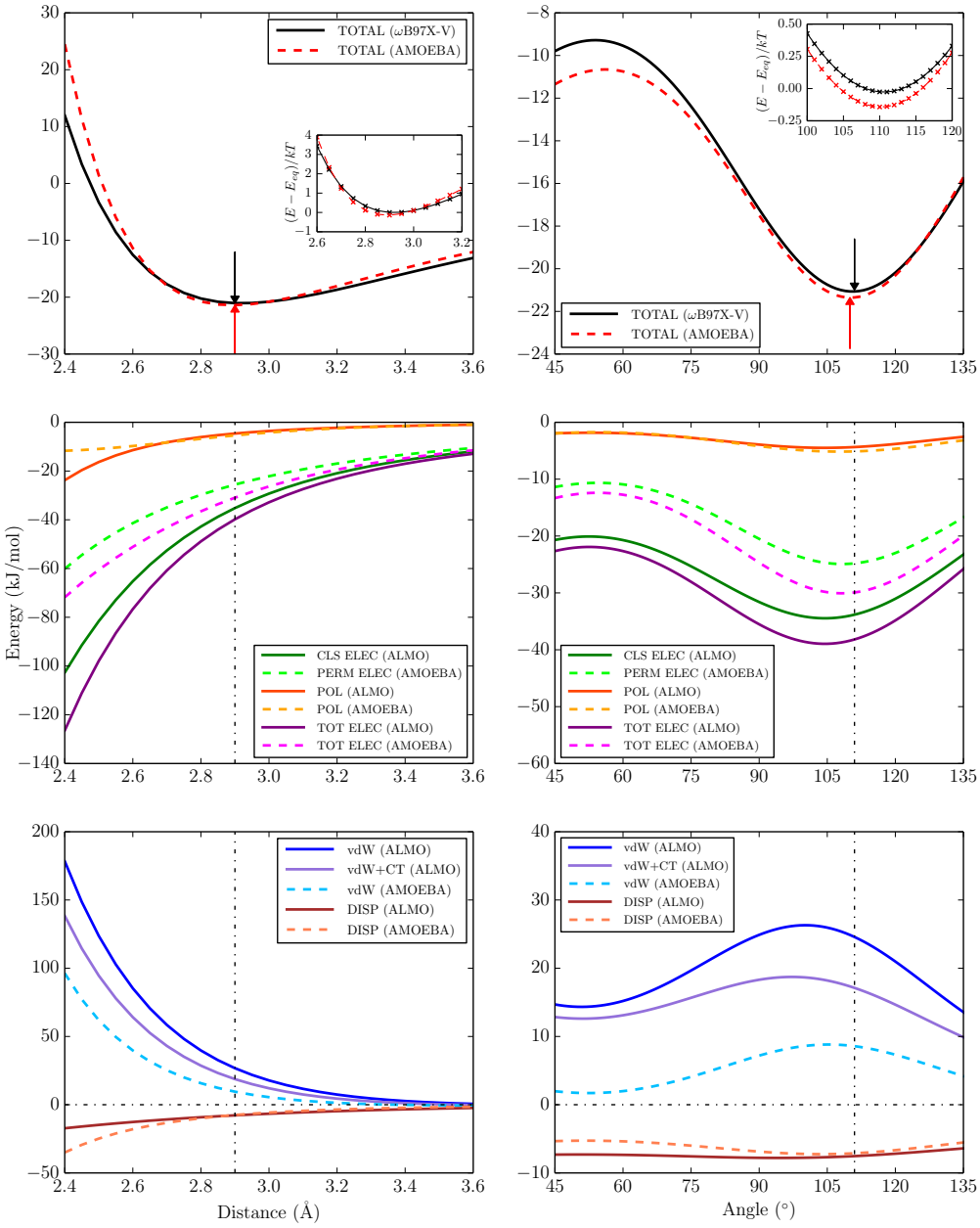

Figure 5.1: Distance (left) and angular (right) dependence of the total interaction energy and its breakdowns (in kJ/mol) for the water dimer. Top: total interaction energy; middle: permanent and induced electrostatics; bottom: vdW interaction. The inset plots in the two top panels show the zoomed-in near-equilibrium region in the units of kT, and the arrows indicate the location of energy minima for QM and AMOEBA interactions, while the dash-dotted lines in the lower four panels indicate the position of QM minimum.

discrepancy in permanent electrostatics.

From the bottom left plot, we see a very large discrepancy between AMOEBA's total vdW interaction and the physically pertinent terms in ALMO-EDA (modified Pauli + dispersion): AMOEBA's total vdW interaction is more favorable by about 17 kJ/mol at equilibrium, and the difference becomes much more pronounced at shorter separations. Since the attractive component of AMOEBA's vdW agrees with the dispersion term given by ALMO-EDA for most distances (except for the highly compressed region), the large difference in their total vdW interaction must reside in the repulsive part of AMOEBA's 14-7 potential which turns out to be excessively soft. This may seem surprising given the overly repulsive wall of AMOEBA's total interaction energy profile. However, the comparison to EDA shows that the softened repulsive vdW potential of AMOEBA is accounting for two attractive contributions that are important in the short range but which are not explicitly included in AMOEBA: (i) the effect of CP, which renders the QM permanent electrostatics more favorable, and (ii) the CT from the proton acceptor to the proton donor, as an extra stabilizing effect. If we combine the CT term with the total vdW interaction given by ALMO-EDA, it almost halves the gap between AMOEBA and ALMO-EDA's vdW profiles, as the difference at equilibrium reduces to 9.5 kJ/mol. Thus, the energetically favorable CP and CT contributions seem to be implicitly accounted for via the softened vdW potential in AMOEBA, which results in the cancellation of errors that yields good agreement in total interaction energies. However, in the short range, the cancellation of errors turns out to be imperfect, and the extra hardness of AMOEBA's repulsive wall actually resides in its too unfavorable permanent electrostatics and polarization.

It is often deemed to be an important and challenging task for a force field to correctly reproduce the directionality of hydrogen bonds [285, 308, 393]. Therefore, an assessment of the angular dependence of the water dimer interaction will be instructive. The angular scan is performed by modifying the θ angle illustrated in Figure 5.2 at the equilibrium O$\cdots$O distance (2.92 Å) with all the other degrees of freedom fixed. The results for $\theta = 45°$–$135°$ are plotted in the right three panels. According to the ALMO-EDA results, the directionality of the hydrogen bonding interaction in the water dimer is a consequence of the interplay of permanent electrostatics, Pauli repulsion, and CT, while the angular dependence of polarization and dispersion is less appreciable. AMOEBA reproduces the angular dependence predicted by ωB97X-V fairly well in the favorable region: the energy minimum appears at 110°, which is only minimally different from the QM result (111°), and the energy discrepancy is less than 0.3 kJ/mol in the entire low-energy region of the potential well (100-120°, see the inset plot in the top right panel). In the "more exotic" higher energy region ($< 70°$), the AMOEBA curve is slightly too favorable by 1.0–1.5 kJ/mol.

Nevertheless, when turning to the energy breakdowns, we see a sharp difference between QM and AMOEBA's permanent electrostatics, while their polarization profiles exhibit very good agreement in general, which are consistent with the trend observed in the distance scan. As a result, AMOEBA's total electrostatics is too unfavorable by a considerable amount compared to the ALMO-EDA results (the maximum discrepancy is about 11 kJ/mol at $\theta \approx 80°$), and the angle that corresponds to most favorable total electrostatics is shifted

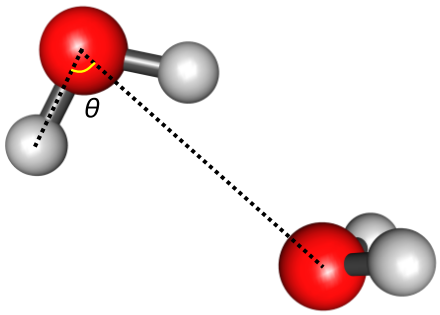

Figure 5.2: Illustration of the angular scan performed for the water dimer: θ is the angle being modified, while the O$\cdots$O distance remains unchanged.

from 104° (the QM value) to 108°. We expect that such a discrepancy is offset by AMOEBA's vdW interaction in order to restore the correct angular dependence. This is confirmed by the bottom right plot: the 14-7 potential of AMOEBA is less repulsive by roughly the same amount when measured against the vdW+CT contribution determined by ALMO-EDA, and the maximum of this net repulsive energetic contribution is also shifted from 97° to 105°.

5.3.2 Water-monovalent ions

The same analysis is then performed on water interacting with Na^+ and Cl^-, as two representative monovalent ions. Compared to the water dimer case where both fragments are neutral, we expect to see much stronger permanent electrostatic interactions (led by charge-dipole interaction) and polarization effects in these systems. The total interaction energies calculated by ωB97X-V and AMOEBA upon rigid dissociation of the $H_2O\cdots Na^+$ complex (conserving C_{2v} symmetry) are shown in the top left panel of Figure 5.3. We see very good agreement between them (even in the repulsive region), although we should bear in mind that the energy scale here (increment of the y axis) is much larger than that in the water dimer case. The O$\cdots$$Na^+$ distance corresponding to AMOEBA's energy minimum (2.25 Å) matches the QM value (2.20 Å) closely, while AMOEBA slightly underbinds the complex in the whole plotted range (by about 4 kJ/mol in the vicinity of equilibrium).

The magnitude of permanent electrostatics for this system is over 100 kJ/mol at the equilibrium separation. Due to the lack of CP in AMOEBA, its permanent electrostatics is consistently less favorable than its ALMO-EDA counterpart. The discrepancy at equilibrium is roughly 8.5 kJ/mol. They start to match to within 1 kJ/mol not far beyond that (> 2.75 Å), which validates the accuracy of the distributed-multipole scheme in describing permanent electrostatics in the long range.

The discrepancy between polarization energies evaluated by QM and AMOEBA varies with intermolecular separation. While AMOEBA only slightly overestimates the polarization energy (by about 3.5 kJ/mol) at equilibrium, it significantly overpolarizes in the short range (almost twice as favorable as ALMO-EDA's polarization in the more compressed region).

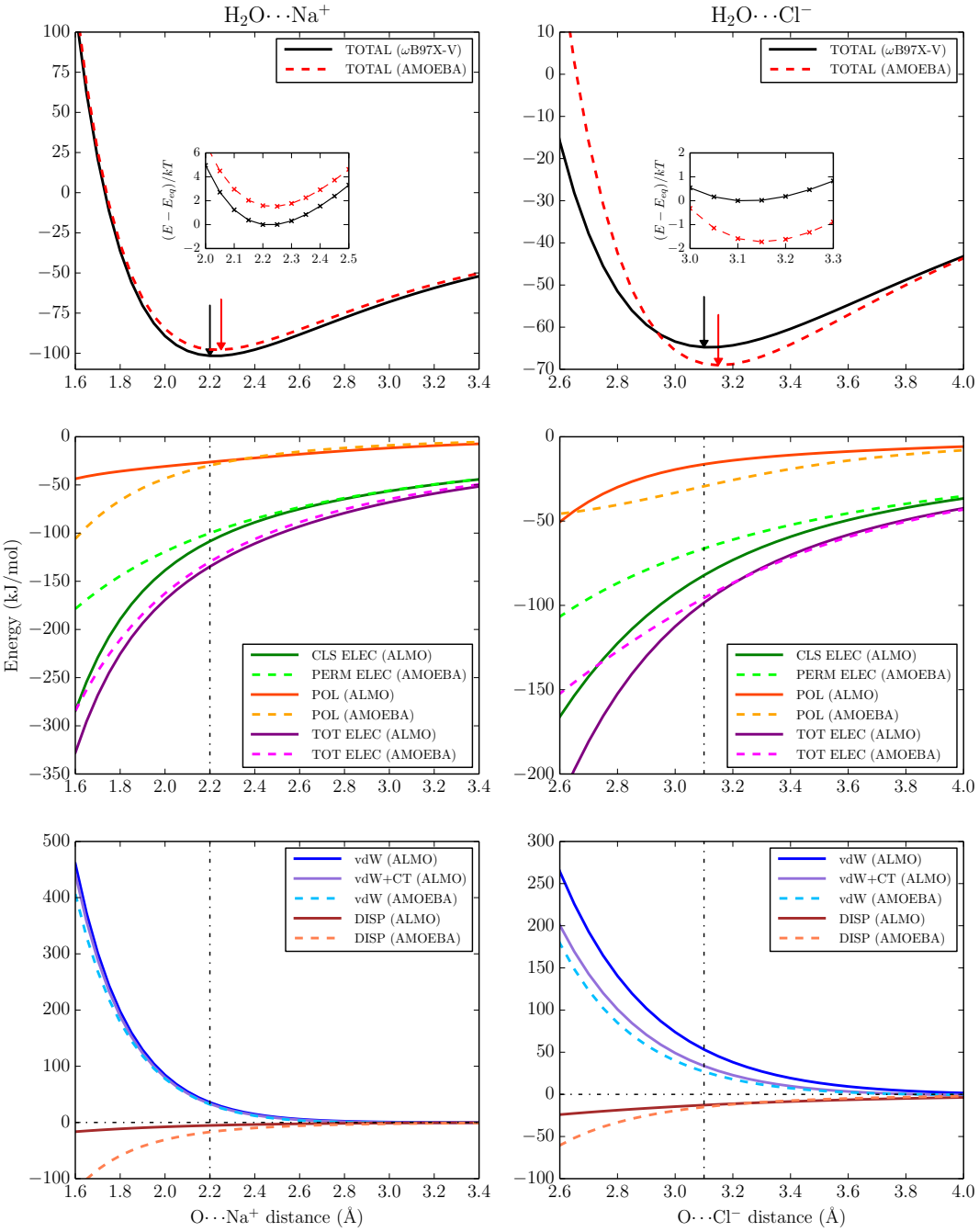

Figure 5.3: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid dissociation of $H_2O\cdots Na^+$ and $H_2O\cdots Cl^-$ complexes. The plotting details (arrangement of figures and symbols used) are the same as in Figure 5.1.

This suggests inadequate damping from the Thole model. Nevertheless, at short range, AMOEBA's overestimated polarization appears to balance its too unfavorable permanent electrostatics, which results in reasonable agreement with QM for the total electrostatic contributions.

The almost superimposed curves for AMOEBA and ALMO-EDA's vdW interactions (see the bottom left panel) indicates their close agreement, despite the sharp difference seen for their attractive component (dispersion). Also, the contribution from CT is almost negligible for this system compared to the magnitude of the total interaction energy.

The performance of AMOEBA is also assessed on the $H_2O\cdots Cl^-$ complex, and the results are shown in the right three panels of Figure 5.3. While comparison of the two top panels suggests a more considerable difference between QM and AMOEBA's total interaction energy profiles, it should be kept in mind that the energy range plotted for this system is much smaller. In fact, the performance of AMOEBA (size of errors) in the vicinity of equilibrium is close to that in the $H_2O\cdots Na^+$ case: it shifts the energy minimum to longer distance by 0.05Å, and overestimates the equilibrium interaction energy by 4-5 kJ/mol. However, the total interaction energies given by QM and AMOEBA start to differ significantly when entering the compressed region. AMOEBA becomes repulsive more rapidly for $R_{O\cdots Cl^-} < 2.9$ Å, although it is overbound adjacent to and beyond equilibrium. Also, for $R_{O\cdots Cl^-} > 4.2$ Å (not shown on the plot), AMOEBA's total interaction energy again becomes less favorable than the QM result.

One might ascribe AMOEBA's much harder repulsive wall to an improperly trained 14-7 potential. However, according to the bottom right panel of Figure 5.3, the 14-7 potential of AMOEBA is less repulsive than ALMO-EDA's vdW contribution across the entire plotted range. Its fairly reasonable agreement (differing by ∼7 kJ/mol at equilibrium) with ALMO-EDA's vdW+CT term indicates again that AMOEBA implicitly incorporates CT through its 14-7 potential.

Therefore, similar to the water dimer case, the deficiency of AMOEBA for the short-range interaction between H_2O and Cl^- mostly arises from the failure to fully compensate for the missing effect of CP via the 14-7 potential. The middle right plot of Figure 5.3 shows that AMOEBA's permanent electrostatic interaction is significantly underestimated (by about 15 kJ/mol) in the vicinity of the energy minimum. This energetic discrepancy, nonetheless, is largely canceled by its overestimated polarization energy in the same range, which leads to fairly close agreement in total electrostatics between QM and AMOEBA at and beyond equilibrium. The error at the bottom of AMOEBA's total interaction potential well is mostly due to the aforementioned slight difference (whose size is only a few kJ/mol) between the 14-7 potential of AMOEBA and ALMO-EDA's vdW+CT contribution. However, at shorter intermolecular separations, this small discrepancy in vdW interactions is overwhelmed by the error due to AMOEBA's far too unfavorable permanent electrostatics. At the same time, AMOEBA's overestimation of polarization also diminishes gradually with reduced intermolecular distance, due to the onset of Thole damping, which leaves permanent electrostatics as the culprit for the excessively hard repulsive wall in the total interaction potential of $H_2O\cdots Cl^-$.

Comparing the results of permanent electrostatics in $H_2O\cdots Na^+$ and $H_2O\cdots Cl^-$ at their individual equilibrium separations, AMOEBA underestimates this energetic component by 8% and 19%, respectively, due to the lack of explicit account of CP. It is not surprising that the latter system is more prone to CP, since the charge distribution of Cl^- is much more diffuse than that of Na^+.

The difference between AMOEBA's polarization profiles for these two systems at short range is also intriguing, and must arise from the effect of Thole damping through the effectively smeared point multipoles. Based on Eq. (5.8), atomic sites with higher polarizability are more smeared in the damping procedure. The polarizability value for Cl^- is 4.00 Å^3, which is one of the largest amongst all atom types available in the AMOEBA force field. For $H_2O\cdots Cl^-$, although AMOEBA's polarization energy is more favorable than ALMO-EDA's over almost the entire plotted range, the onset of Thole damping is clearly manifested in the curvature in the AMOEBA polarization profile. Indeed there is an inflection point near the equilibrium distance, and therefore AMOEBA polarization crosses with ALMO-EDA's polarization curve at $R_{O\cdots Cl^-} = 2.65$ Å. In contrast, the onset of damping is not apparent in $H_2O\cdots Na^+$, such that the polarization effect due to the cation is significantly overestimated by AMOEBA in the short range. This is due to the considerably smaller atomic polarizability of Na^+ (0.12 Å^3), rendering the damping effect through Eq. (5.8) negligible unless Na^+ and H_2O are in extremely close contact.

5.3.3 Water-divalent cations

It is well-known that the description of water-divalent cation interactions is challenging for classical force fields[135, 418]. We next assess the agreement between AMOEBA and ωB97X-V results upon rigid dissociation of $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ complexes (with C_{2v} symmetry). The results are collected in Figure 5.4. For both systems, AMOEBA correctly reproduces the position of energy minima, while it underestimates the magnitude of the binding energies across the board. At the equilibrium distance, AMOEBA underbinds $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ by 21 kJ/mol and 17 kJ/mol, respectively, which is 6–7% of the total interaction energies evaluated by ωB97X-V.

To understand the much larger deviations compared to that in the $H_2O\cdots Na^+$ case, we again measure the individual terms of AMOEBA against the energy components given by ALMO-EDA. Due to the +2 charge, permanent and induced electrostatics dominate the strongly favorable total interactions. Some qualitative similarities emerge for these two systems, as seen in the middle two panels of Figure 5.4. First is the consistently less favorable short-range permanent electrostatics of AMOEBA, which is a common issue for all the systems assessed so far. Second is the crossing of the AMOEBA and ALMO-EDA polarization energy curves slightly beyond the equilibrium distance.

The permanent electrostatic interactions in these two systems are *identically* described by the AMOEBA model (+2 point monopole for both Mg^{2+} and Ca^{2+}). However, the short-range discrepancy from ALMO-EDA's classical electrostatics is larger for $H_2O\cdots Ca^{2+}$, indicating a more pronounced CP effect. The CP effect (regarded as the difference between

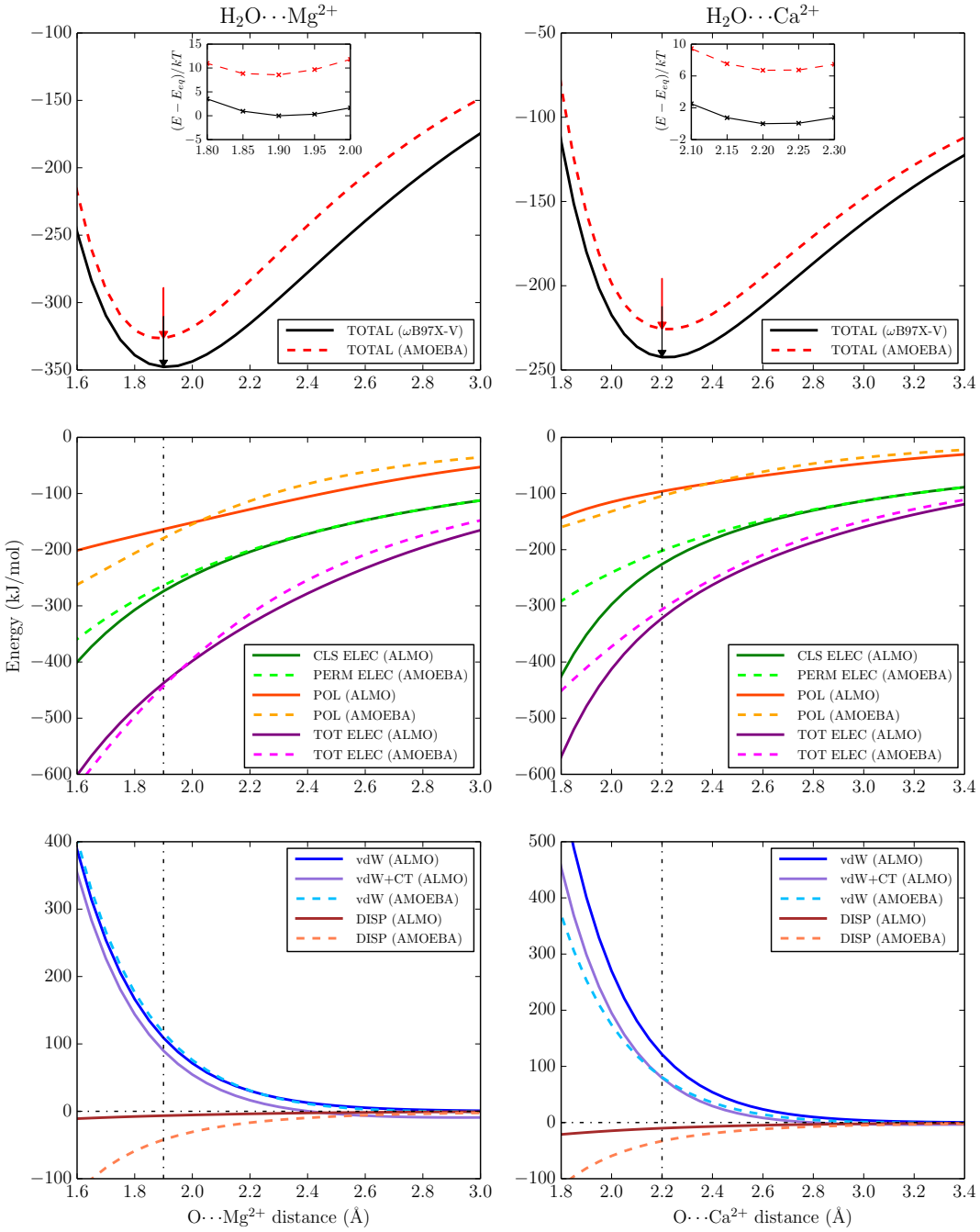

Figure 5.4: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid dissociation of $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ complexes. The plotting details (arrangement of figures and symbols used) are the same as in Figure 5.1.

these two models) is determined by (i) the extent of charge distribution on each fragment and (ii) the intermolecular distance. The former should be the dominant factor here, since otherwise we would expect the more significant CP effect in $H_2O\cdots Mg^{2+}$ whose equilibrium distance is 0.3 Å shorter. This trend (and interpretation) also applies to the interactions between water and alkali metal cations (results for $H_2O\cdots Li^+$ and $H_2O\cdots K^+$ are shown in Figure D.2), where the effect of CP increases as $Li^+ < Na^+ < K^+$.

The more diffuse charge distribution in $H_2O\cdots Ca^{2+}$ is also be reflected in the (modified) Pauli term in ALMO-EDA. At equilibrium, the Pauli term for $H_2O\cdots Ca^{2+}$ is 16 kJ/mol more repulsive than that for $H_2O\cdots Mg^{2+}$ despite the 0.3 Å longer intermolecular separation of the former, indicating stronger overlap between fragment charge densities resulting from the much more diffuse charge distribution of Ca^{2+}.

The polarization energies given by AMOEBA for these two systems, at their equilibrium intermolecular separations, are slightly overestimated (by roughly 9%). This order is reversed (AMOEBA polarization becomes too weak) in the long range. In the compressed region plotted in Figure 5.4, AMOEBA overpolarizes in both cases. However, one nuanced distinction exists: for $H_2O\cdots Mg^{2+}$, the amount of AMOEBA's over-polarization increases monotonically with the shortening of $O\cdots Mg^{2+}$ distance, while in the $H_2O\cdots Ca^{2+}$ case, the difference reaches a maximum at 1.90 Å (18.5 kJ/mol), and a crossing point emerges at more compressed distance (1.65 Å, not shown in the figure). This behavior is similar to what was discussed previously for $H_2O\cdots Cl^-$, which was regarded as a signature of the onset of Thole damping. Here it can also be explained by the same polarizability-based argument. The AMOEBA atomic polarizability for Mg^{2+} (0.08 $Å^3$) is so small that the effect of damping can hardly be seen in the entire plotted range. By contrast, the value for Ca^{2+} is significantly larger (0.55 $Å^3$) so that the damping effect is manifested in the slightly compressed region. The same trend is observed with the polarization energy curves of water-alkali metal cation series (the results for Li^+ and K^+ are shown in Figure D.2), where AMOEBA immensely overestimates the polarization energy of $H_2O\cdots Li^+$ in the short range, while the agreement with ALMO-EDA's polarization is much better for $H_2O\cdots K^+$.

In terms of total electrostatics, AMOEBA agrees closely with ALMO-EDA near the equilibrium distance in the $H_2O\cdots Mg^{2+}$ case (the former is more favorable by 5 kJ/mol at equilibrium), thanks to error cancellation between permanent electrostatics and polarization. The deviations in the short range and long range have opposite signs, which, in contrast to many other systems, are both dominated by the discrepancy in polarization. For $H_2O\cdots Ca^{2+}$, AMOEBA's total electrostatics is less favorable across the entire range, and the discrepancy at equilibrium is about 3 times larger than that in the $H_2O\cdots Mg^{2+}$ case, as AMOEBA underestimates the permanent electrostatic interaction more severely in $H_2O\cdots Ca^{2+}$ as a consequence of the increased importance of CP.

Turning to the vdW terms, a remarkable common feature of the bottom two panels of Figure 5.4 is the much more attractive $1/R^7$ potential in AMOEBA relative to the dispersion component determined by ALMO-EDA. Specifically, at the equilibrium distance of each system, AMOEBA's dispersion (attractive vdW potential) is 6.5 times more favorable in $H_2O\cdots Mg^{2+}$, and 3 times more favorable in $H_2O\cdots Ca^{2+}$. It is inevitable for the $1/R^7$

potential of AMOEBA to be excessively attractive in the strongly overlapping regime since it is not appropriately damped, however, we think that it should be able to approximately match the dispersion energy given by ALMO-EDA near the equilibrium region. Here (and for $H_2O\cdots Li^+$ and $H_2O\cdots Na^+$ as well) the large difference against ALMO-EDA's dispersion term could be explained by the incorporation of other stabilizing effects like CP or CT in AMOEBA's $1/R^7$ potential, as we discussed before. On the other hand, it is possible that such differences are related to the fact that AMOEBA's 14-7 potential is parameterized simultaneously so that its repulsive and attractive components may not correspond to their presumed physical meanings when scrutinized individually.

For the total vdW interaction between H_2O and Mg^{2+}, AMOEBA exhibits a slightly more repulsive potential than its ALMO-EDA counterpart when $R_{O\cdots Mg^{2+}} < 2.2$ Å. The inclusion of CT further enlarges the discrepancy, as opposed to the general trend observed in other systems. The plot for $H_2O\cdots Ca^{2+}$, on the other hand, demonstrates the common trend, where AMOEBA's total 14-7 potential is considerably softened and matches ALMO-EDA's vdW+CT curve fairly well except in the strongly unfavorable region. It should be noted that the contribution of CT is appreciable when water interacts with these divalent cations, especially for $H_2O\cdots Ca^{2+}$ where the CT energy at equilibrium is -42 kJ/mol (roughly 1/6 of the total interaction energy). According to our results, this significant stabilizing energy component is implicitly incorporated in the softened 14-7 potential of AMOEBA.

So is there an overall reason for why AMOEBA underbinds these two systems? We suggest, based on the assessment against the ALMO-EDA results, that for $H_2O\cdots Mg^{2+}$ the error mostly resides in AMOEBA's overly repulsive vdW potential. On the other hand, AMOEBA's too unfavorable permanent electrostatics (due to the lack of CP) is the main origin of the underestimated interaction energy between H_2O and Ca^{2+}. Therefore, although these two divalent cations belong to the same main group on the periodic table, and AMOEBA manifests rather similar systematic errors on them, we think that they actually arise for different reasons.

5.3.4 Other water-halide interactions

Although AMOEBA underestimates the total $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ interaction energies, it gives a reasonable description of the shape of their PESs, which is rather important for accurately computing the intermolecular forces. In contrast, the shape of AMOEBA's total interaction energy profile for $H_2O\cdots Cl^-$ agrees less satisfactorily with that generated by QM. Therefore, we complete the full series of water-halide interactions by assessing the performance of AMOEBA on $H_2O\cdots F^-$ and $H_2O\cdots Br^-$ against QM, and the results are shown in Figure 5.5.

As for $H_2O\cdots Cl^-$, AMOEBA gives overly repulsive walls for water-halide interactions in the short range, and also exhibits overly long intermolecular distances. The discrepancy between QM and AMOEBA is most exaggerated in the $H_2O\cdots F^-$ case, where the equilibrium $O\cdots F^-$ distance given by AMOEBA is 0.25 Å too long. At the distance corresponding to the QM minimum ($R_{O\cdots F^-} = 2.45$ Å), AMOEBA is underbound by about 35 kJ/mol. On

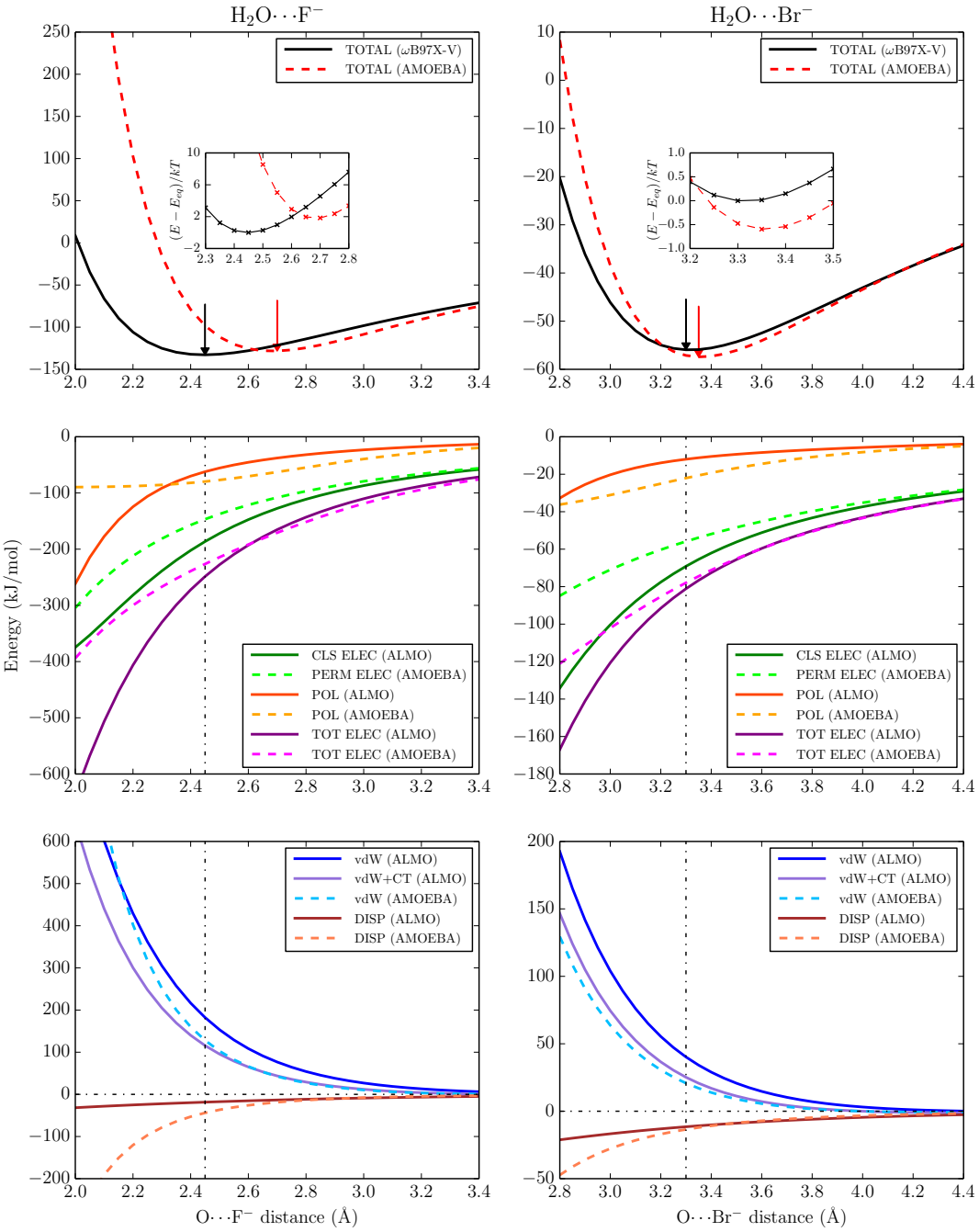

Figure 5.5: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid dissociation of $H_2O\cdots F^-$ and $H_2O\cdots Br^-$ complexes. The plotting details (arrangement of figures and symbols) are the same as in Figure 5.1.

the other hand, the total interaction energy profile of AMOEBA for $H_2O\cdots Br^-$ exhibits features that are rather similar to the $H_2O\cdots Cl^-$ case: the equilibrium distance is slightly overestimated by 0.05 Å, and the energy discrepancy at $R_{O\cdots Br^-} = 3.3$ Å (the QM minimum) is rather small (about 1 kJ/mol).

Based on the ALMO-EDA results, we investigate the reason for such a tremendous discrepancy between QM and AMOEBA's total interaction energy profiles for $H_2O\cdots F^-$. According to the middle left panel of Figure 5.5, at the QM minimum, the permanent electrostatic interaction given by AMOEBA is less favorable than ALMO-EDA's classical electrostatics by over 20% (40 kJ/mol). Due to the diffuse charge distribution of F^- and the relatively small equilibrium distance, the effect of CP should be the main origin of such a difference. Nevertheless, AMOEBA's too unfavorable permanent electrostatics here is not qualitatively different from the situations in other water-halide interactions: e.g., for $H_2O\cdots Cl^-$, AMOEBA underestimates the permanent electrostatic interaction by roughly 20% as well at equilibrium and by over 1/3 in the most compressed region. Therefore, the neglect of CP should *not* be the only culprit for AMOEBA's poor performance on $H_2O\cdots F^-$ *overall*.

As in $H_2O\cdots Cl^-$, AMOEBA's polarization energy is more favorable than that given by ALMO-EDA at and beyond the QM minimum. However, it seems to be damped too quickly after entering the compressed region. Specifically, the AMOEBA polarization curve is almost flat when $R_{O\cdots F^-} < 2.45$ Å, and it becomes less favorable than ALMO-EDA's polarization energy when $R_{O\cdots F^-} \leq 2.3$ Å. The overdamping of AMOEBA's polarization at short range is essentially similar to what we observed in other water-halide interactions, but it is far more pronounced in this system, most likely due to the much shorter interfragment distance. Therefore, the permanent and induced electrostatics of AMOEBA are *both* too unfavorable in the short range, contributing to the striking difference between AMOEBA and ALMO-EDA's total electrostatic contributions together. It is also noteworthy that AMOEBA's total electrostatics turns out to be more favorable than its ALMO-EDA counterpart when $R_{O\cdots F^-} > 2.6$ Å (the largest difference beyond that point is about 6 kJ/mol), where the less attractive permanent electrostatics of AMOEBA is outweighed by its more favorable polarization energy, which is also observed in $H_2O\cdots Cl^-$ and $H_2O\cdots Br^-$ systems (though to a lesser extent).

In the previous discussion of the $H_2O\cdots Cl^-$ complex, we ascribed the overly repulsive AMOEBA potential wall to its too unfavorable permanent electrostatics. While AMOEBA's vdW interaction appears to be softer than ALMO-EDA's vdW+CT contribution, it partially compensates for this deficiency in electrostatics. According to the two bottom panels of Figure 5.5, a similar trend emerges for $H_2O\cdots Br^-$, and it again leads to the reasonable agreement between QM and AMOEBA's total interaction energies, at least near the bottom of the potential well.

However, the situation is rather different for $H_2O\cdots F^-$. While the vdW potential of AMOEBA matches ALMO-EDA's vdW + CT contribution closely at and beyond the QM equilibrium, its repulsive part rises too rapidly at compressed $O\cdots F^-$ distances. For instance, at 2.2 Å (0.25 Å shorter than the (QM) equilibrium distance), the vdW interaction

of AMOEBA is already 70 kJ/mol more unfavorable than the ALMO-EDA's vdW+CT contribution. Therefore, based on the analysis above, AMOEBA's total electrostatic and vdW interactions are both too unfavorable in the short range. These two errors accumulate, rather than cancel each other, which result in AMOEBA's very poor description of the total PES for the $H_2O\cdots F^-$ complex.

5.4 Discussion

In comparing the breakdown of AMOEBA's classical molecular interactions against the energy decomposed QM results for the water-water and the water-monovalent cation interactions, it is apparent that the observed good agreement in total interaction energy rests on a very delicately balanced cancellation of errors, as would be expected for most if not all empirical force fields. In the compressed region, where AMOEBA exhibits polarization that is typically too favorable until the onset of Thole damping is realized, it is overwhelmed by the excessively unfavorable permanent electrostatics that originates in the failure of the point multipole model to account for charge penetration, an important stabilizing effect which becomes more pronounced at short range. The fact that the buffered 14-7 potential contains short-range softening effects to implicitly compensate for the electrostatics is perhaps not surprising given the parameterization procedure of AMOEBA wherein the permanent multipoles and polarization parameters are determined first and all the remaining non-covalent energetic effects are folded into the 14-7 potential [139, 140].

For the water-halide cases where agreement in total interaction energies between QM and AMOEBA are poorer, the cancellation of errors is insufficient for two reasons. First, too much is asked of the limited functional form of the 14-7 vdW potential to account for favorable CP and CT effects at short range, which are both more pronounced for anions than for cations. Second, for halides, the AMOEBA polarization suffers from overdamping in roughly the same region, which is a consequence of the exclusive functional dependency of the Thole damping on the atomic polarizabilities (Eq. (5.8)), whose values are very large for the halides.

For the water-divalent cation systems, the total interaction energies in AMOEBA are significantly underestimated compared to QM throughout the distance range. For $H_2O\cdots Ca^{2+}$, the breakdown of the interaction energy into individual components exhibit the same problems observed for the water-halide systems. However, it may also be exacerbated by the fact that the original fit of AMOEBA's divalent cation parameters was to a somewhat limited QM benchmark, where the counterpoise corrections for BSSE were performed with insufficiently large basis sets [309, 310], which may undershoot the correct interaction energies of these systems. Interestingly, $H_2O\cdots Mg^{2+}$ shows another difference in its 14-7 potential, which is even more repulsive than the sum of ALMO-EDA's Pauli repulsion and dispersion so it evidently cannot implicitly incorporate any CT. In addition, its permanent electrostatics at short range is less unfavorable versus the $H_2O\cdots Ca^{2+}$ case, mostly due to the comparatively more compact Mg^{2+} charge distribution.

Although errors in AMOEBA's polarization contribute less significantly to errors in total interaction energy in general, it clearly deviates from ALMO-EDA's polarization profiles in almost all cases. Taken together, these point to either a shortcoming in the functional form given by Eq. (5.8), or the way in which the damping parameters are determined. In the original Thole model, atomic polarizabilities and the damping parameter were determined by a fit to molecular polarizabilities [410, 411]. However, molecular polarizability is *not* very sensitive to either the value of the damping parameter or even the functional form of the damping, since a linear model performs as well as the currently used exponential model [410]. On the other hand, it has been found that the same "dimensionless width" parameter, a, in Eq. (5.8) cannot simultaneously reproduce the gas-phase molecular polarizability of water and QM energies of small clusters of water molecules, which indicates the higher sensitivity of polarization energetics to the choice of this value [140, 384].

It has been suggested that the need to modify the damping parameter, is a manifestation of exchange-polarization coupling between molecules that occurs in the regime of overlapping charge distribution.[328, 419, 420] Such an effect is intrinsically accounted for in ALMO-EDA, since its polarization term is computed using a properly antisymmetrized wavefunction. Therefore, the discrepancies between ALMO-EDA and AMOEBA's polarization energies observed in this work may also be related to this missing effect in AMOEBA. It is noteworthy that only a single value of the damping parameter, a, is employed for almost every type of interaction in AMOEBA (except for aromatic carbon atoms and divalent cations [140]), and therefore the damping may benefit from greater chemical specificity in choosing this parameter in order to capture differences in diffuseness of atomic charge distributions, as well as the exchange-polarization coupling effect at short range. Moreover, a similar argument to incorporate greater chemical specificity may also apply to atomic polarizabilities, since by simply reproducing the isolated molecular polarizabilities within the current parameterization, the resulting polarizability values may not yield correct polarization energetics under more complicated interacting environment.

An additional potential limitation of the AMOEBA functional form may reside in its reliance on isotropic polarizabilities, which can be seen in the angular scans for water interacting with alkali metal cations that show larger discrepancies between the AMOEBA and ALMO-EDA polarization energies in the strained region (see Figures D.3 and D.4 in Appendix D), due to the presence of stronger electrostatic fields. This points to two aspects of the anisotropy of polarization. The first is the issue of whether additional polarization sites other than the atom centers are needed to faithfully reproduce properties determined by QM. The polarization profiles of AMOEBA seem to overemphasize the angular dependence, which might be related to the lack of extra inducible sites around the water oxygen. Indeed, it has been shown that MM models that incorporate polarization centers at the lone pair sites of the water oxygen in addition to the atom centers more faithfully reproduce QM-derived energies and dipole moments for several sets of water oligomers than models that use atom-centered polarization only[421].

The second aspect of the anisotropy is the replacement of an isotropic scalar polarizability with a polarizability tensor. Harder et al. examined both of the major aspects of polarization

anisotropy in the context of the Drude oscillator model [422]. They observed that the use of both lone pair polarizability sites and anisotropic polarizabilities is essential to faithfully reproduce the QM electrostatic potentials along a curvilinear coordinate, compared with using isotropic polarizabilities either with or without lone-pair polarizability sites. In terms of the energetics of oligomeric systems of small organic molecules in water clusters, the use of lone pair polarization sites turns out to be the more important aspect, and the additional use of an anisotropic polarizability tensor further reduces the error with respect to the QM results.

5.5 Conclusion

In this work, we have compared the energetic profiles of water-water and water-ion dimers generated by an advanced MM force field, AMOEBA, and those obtained with the ωB97X-V density functional whose accuracy was verified by a comparison to the highly accurate ΔCCSD(T)/CBS reference values. More importantly, we have appraised the force field by comparing the relative contributions of its non-covalent terms with corresponding terms generated by a decomposition of the DFT total interaction energy using the second generation of the ALMO-EDA.

Taken together, a number of trends have emerged. It is clear that the physical effects that are missing in AMOEBA, specifically charge penetration (CP) and charge transfer (CT), have been captured implicitly in the 14-7 vdW potential. This is a consequence of both the short-range buffering that perhaps renders it amenable to capturing such short-range softening (stabilizing) effects, and the fact that the vdW parameters are determined at the end of the parameterization of the non-covalent terms in AMOEBA to match the binding energies given by QM reference. However, it is also clear that this implicit accounting of short-ranged softening effects by the 14-7 potential is imperfect, as revealed by the investigation on water-divalent cation and water-halide interactions. In addition, the Thole-damping of AMOEBA's polarization was sometimes found to yield unphysical results at short range (underdamped for water-cations while overdamped for water-anions, in general), which is related to the exclusive dependence of the damping effect on atomic polarizabilities. Perhaps the ALMO-EDA (or related methods) can help refine a next generation AMOEBA model that realizes a better cancellation of errors for the problematic ion-water cases.

Alternatively, the ALMO-EDA could also be used to guide the development of explicit functional forms and the associated parameterization for short-range effects like CP and CT. Indeed, efforts are underway by others to explicitly account for the effect of CP under the framework of the AMOEBA force field[142, 367], and simply adding corrections for CP to the monopole-monopole term in AMOEBA's permanent electrostatics has been shown to yield considerably improved agreement with the permanent electrostatics given by the SAPT2+ level of theory.[142] Nonetheless, to yield a balanced force field, AMOEBA's 14-7 potential would need to be reparameterized in the context of explicit inclusion of CP to avoid over-counting this effect. Similarly, any future effort to incorporate CT explicitly would require

reparameterization of the 14-7 potential as well. The separation of these short-range softening effects from the vdW potential has benefits. In particular, its repulsive and attractive components could be parameterized individually according to their accepted physical meaning (i.e., Pauli repulsion and dispersion, respectively) which have precise definitions within QM methods such as the ALMO-EDA.

Chapter 6

QM/AMOEBA: Formulation and Assessment

6.1 Introduction

The solvation process, where a solute molecule interacts with water or other solvent molecules, plays a key role in many chemical and biochemical systems. In chemical reactions, the success with which inorganic or organic compounds are synthesized can be greatly affected by the choice of the solvent. In macromolecular systems, the rates for enzymatic reactions or protein folding and the strengths of ligand-receptor binding are also dependent upon solvation, because all these processes involve partial exposure of key chemical groups to the solvent.

Given its central role in both basic and applied sciences, molecular solvation has been investigated using *ab initio* quantum chemistry methods in conjunction with implicit or explicit solvent models in numerous theoretical and computational studies. Implicit solvent models [423], which ignore the molecular resolution of the solvent, have a long history dating back to Born [424] and Onsager [425]. In quantum mechanical (QM) calculations, implicit descriptions for solvent molecules remain in wide use today, mostly in the formulation of generalized Born (GB) models [426–428], apparent surface charge (ASC) models [429–440], and models based on direct solution of non-homogeneous Poisson-Boltzmann equations [441–444]. Despite their huge successes — solvation free energies for neutral molecules can be predicted on average within 1.0 kcal/mol [445, 446]— implicit solvent models for QM calculations can be inadequate in several situations: (a) Larger errors in the predicted solvation free energies are found for ionic solutes [445], which can interact strongly with solvent molecules through permanent electrostatics, polarization, and charge transfer; (b) It is rather difficult to describe systems that are partly inhomogeneous, such as ionic liquids (and other mixed solvents) or one solvent at different pH conditions; (c) It is clearly ill-suited for describing completely inhomogeneous environments, such as a "solute" ligand in a partially exposed binding pocket that is accessible to water or other solvent molecules.

Explicit solvent models, where the solvent molecules receive an all-atom or united-atom description, can in principle be employed in such situations that are challenging for implicit solvent models. Ideally, one would like to perform fully *ab initio* QM (such as density functional theory (DFT) or perturbation theory (PT)) calculations on the solute molecule together with a larger number of solvent molecules. While there have been many fully *ab initio* molecular dynamics (AIMD) or Monte Carlo (MC) simulations on liquid water [447–453] and ion-water systems [454–461] reported, their routine use is still beyond the reach of most researchers, due to the substantial cost associated with computing the *ab initio* electronic structure for each configuration, and the enormous number of configurations required to adequately sample the configuration space. Even when feasible, AIMD simulations using DFT are *not always* guaranteed to produce accurate results for bulk liquid or solute-solvent systems as the quality of results depends heavily on the delicate interplay between the density functional and the dispersion correction employed, [360, 462–468] while many PT methods are known to overestimate dispersion even for smaller systems.

Explicit solvent models utilizing hybrid quantum mechanical molecular mechanical (QM/MM) energy functions [147–152] come as a natural compromise between computational efficiency and accuracy. Through treating the solute molecule with *ab initio* QM methods and the solvent molecules with molecular mechanics (MM) force fields [131–133, 469–472], it vastly reduces the computational demand compared to full QM calculations, while potentially retaining the QM accuracy for the simulation results. Out of three categories of interactions existing in a solute-solvent system (intramolecular interactions within the solute, solute-solvent interactions, and solvent-solvent interactions), *it is the most crucial to have an accurate description for the solute-solvent interactions.* This is because one can usually find a QM level of theory (such as Kohn-Sham (KS)-DFT [5]) to reliably describe intramolecular interactions within the solute, and because, for solvent-solvent interactions, one can take advantage of the error cancellation within the MM model or reduce the error by averaging over the sampling ensemble.

In QM/MM based solvation models, the solute-solvent interactions consist of five components: permanent electrostatics, forward (MM $\rightarrow$ QM) polarization, backward (QM $\rightarrow$ MM) polarization, exchange repulsion, and dispersion. (Note: forward and backward polarizations are numerically inseparable if mutual polarization is enabled.) Most of the recent algorithm developments on QM/MM interactions have focused on the following aspects: a) improving the permanent electrostatics and the forward polarization by adopting a multipolar representation of solvent electrons [153, 154, 168, 473–478] and by introducing damping schemes to account for the spread of solvent valence electron density [153, 154, 479–484]; and b) adding the backward polarization through employing a polarizable force field for the solvent molecules, including models based on Drude oscillators [485–488], fluctuating charges [489–491] and inducible dipoles. [153, 154, 168, 473–478, 492–502]

In QM/MM calculations, exchange repulsion and dispersion interactions are usually combined together in a classical vdW potential using the Lennard-Jones ("12-6") or Halgren ("14-7") formula [412]. To date, there is no fully automated procedure to assign vdW parameters to QM atoms, and there is a lack of systematic ways to improve the description

of QM/MM vdW interactions, which potentially limits the accuracy of QM/MM results in general. But vdW interactions are just as important as permanent electrostatics and polarization interactions in QM/MM calculations [503]. In two recent publications on hydration free energies [504, 505], for example, it was reported that the BLYP functional [34, 35] (for the QM region) can produce more accurate results than more sophisticated functionals, when the solvent water molecules are described by the TIP3P [506] model. This happens only because the QM/MM permanent electrostatics based on BLYP electron density is the most compatible with the employed classical QM/MM vdW potential. In another recent publication [168], we also demonstrated that the buffered 14-7 potential of AMOEBA needs to be adjusted to reproduce full QM values for the investigated solute-solvent interaction energies.

In this article, a new implementation of DFT/AMOEBA calculations within the Q-Chem/LibEFP [255, 507, 508] software framework is reported, with a complete derivation of equations for self-consistent field (SCF) calculations. The AMOEBA polarizable force field, [139–141] which places permanent multipoles (up to quadrupoles) and induced dipoles on each solvent atom, is supposed to improve the description of QM/MM permanent electrostatics and forward/backward polarization interactions. Our implementation complements recent efforts within the Gaussian[478], LICHEM[477], and ONETEP/TINKER [168] software platforms, and further improves the accessibility to mutually polarizable QM/MM calculations using the AMOEBA force field.

Equally importantly, we propose a scheme to decompose the solute-solvent interaction energy evaluated by mutually polarizable QM/MM into contributions from permanent electrostatics, polarization and vdW interaction. This method has similar objectives to the semi-empirical QM/MM energy decomposition analysis (EDA) scheme suggested by Gao [509] that was utilized to elucidate the role of polarization in solute-solvent interactions, and more recently, a similar approach (but not exactly the same) was also proposed by Hirao et al. [510] to investigate the stabilization effect of the protein environment on the active site in a polarizable QM/MM calculation for cytochrome P450cam, which turns out to be a useful tool for interpreting the results of QM/MM protein modeling. Here, we employ the EDA method to facilitate a component-wise comparison against full QM references obtained via performing the absolutely localized molecular orbitals based energy decomposition analysis (ALMO-EDA). [97, 115, 118] This allows us to thoroughly analyze all individual components of the solute-solvent interaction energy, and provides valuable insights for guiding future improvements to the QM/MM modeling of solute-solvent interactions. We note that there have been many other protocols developed for decomposing full QM intermolecular interactions [98] so that the energy components are not uniquely defined. Nevertheless, the choice of the ALMO-EDA scheme (with necessary modifications) has been validated by a recent benchmark study of the AMOEBA force field by several of us. [167] More details about the QM/MM and full QM EDA schemes employed in this work are provided in Sec. 6.2.2.

6.2 Methods

6.2.1 Fully polarizable QM/AMOEBA SCF calculations

The total energy of the coupled KS-DFT/AMOEBA system can be expressed as

$$\begin{aligned} E_{\text{tot}}^{\text{QM/MM}} &= E_{\text{nuc-nuc}}^{\text{QM}} + E_{\text{val}}^{\text{MM}} + E_{\text{elec}}^{\text{MM}} + E_{\text{vdw}}^{\text{MM}} \\ &+ E_{\text{vdw}}^{\text{QM/MM}} + E_{\text{elec}}^{\text{QM(nuc)/MM}} \\ &+ E_{\text{el}}^{\text{QM}} + E_{\text{elec}}^{\text{QM(el)/MM}} + E_{\text{pol}}^{\text{MM}}. \end{aligned} \tag{6.1}$$

The first six terms in Eq. (6.1) do not depend on the electron density of the QM region. Among them, the first four terms can be evaluated with a QM or MM region in isolation: $E_{\text{nuc-nuc}}^{\text{QM}}$ represents the repulsion between QM nuclei, $E_{\text{val}}^{\text{MM}}$ is the sum of all the valence terms (bond, angle, and Urey-Bradley) in AMOEBA, and $E_{\text{elec}}^{\text{MM}}$ and $E_{\text{vdw}}^{\text{MM}}$ refer to the permanent electrostatic and van der Waals (vdW) interactions between AMOEBA fragments, respectively. For a more detailed introduction to these terms, we refer the readers to the original references of AMOEBA. [139–141]

The next two terms are two components of the interaction crossing the QM and MM boundary, i.e., the vdW interaction between QM and MM atoms and the electrostatic interaction between the QM nuclei and the permanent multipoles (up to quadrupoles) of the AMOEBA water molecules. In our current model, the QM/MM vdW interaction energy is computed at the MM level, i.e., we assign AMOEBA's vdW parameters to each QM atom. And for the evaluation of $E_{\text{elec}}^{\text{QM(nuc)/MM}}$, the same equations for computing charge-charge, charge-dipole and charge-quadrupole interactions in AMOEBA can simply be applied.

The last three terms are electron-density-dependent so they need to be minimized through an SCF procedure. $E_{\text{el}}^{\text{QM}}$ is defined as the KS energy associated with the present electron density *without* accounting for its interaction with the MM embedding potential, and we shall denote the corresponding Fock matrix contribution as

$$\mathbf{F}_0 = \partial E_{\text{el}}^{\text{QM}} / \partial \mathbf{P}, \tag{6.2}$$

where $\mathbf{P}$ is the density matrix for the QM electrons.

The second density-dependent term, $E_{\text{elec}}^{\text{QM(el)/MM}}$, refers to the Coulomb interaction energy between the QM electrons and the permanent multipoles on the AMOEBA fragments:

$$E_{\text{elec}}^{\text{QM(el)/MM}} = \int \mathrm{d}\mathbf{r}\ \rho_{\text{el}}(\mathbf{r}) V_{\text{m-pole}}^{\text{MM}}(\mathbf{r}), \tag{6.3}$$

where $V_{\text{m-pole}}^{\text{MM}}(\mathbf{r})$ is the electrostatic potential (ESP) generated by the AMOEBA multipoles. Within the AMOEBA force field, the permanent multipoles ($\{\mathbf{M}_i\}$) are usually located on each atomic site i, and $\mathbf{M}_i = \{q_i, \boldsymbol{\mu}_i, \mathbf{Q}_i\}$. Taking all the atomic sites in the MM region

together, we have

$$V^{\text{MM}}_{\text{m-pole}}(\mathbf{r}) = \sum_{i=1}^{n_{\text{MM}}} \Big(\frac{q_i}{|\mathbf{r}-\mathbf{R}_i|} - \nabla_{\mathbf{r}} \frac{1}{|\mathbf{r}-\mathbf{R}_i|} \cdot \boldsymbol{\mu}_i + \frac{1}{3}\, \nabla_{\mathbf{r}} \nabla_{\mathbf{r}} \frac{1}{|\mathbf{r}-\mathbf{R}_i|} : \mathbf{Q}_i \Big), \tag{6.4}$$

where the point charge, dipole and quadrupole moments on each site are contracted with the electric potential, field and quadrupole field operators, respectively, to give the electrostatic potential at an arbitrary point in the 3-space. Transforming Eq. (6.3) into the atomic orbital (AO) basis, we obtain

$$E^{\text{QM(el)/MM}}_{\text{elec}} = \text{Tr}[\mathbf{P}\mathbf{V}^{\text{MM}}_{\text{m-pole}}] \tag{6.5}$$

and its contribution to the Fock matrix of the coupled QM/AMOEBA system is

$$\mathbf{V}^{\text{MM}}_{\text{m-pole}} = \frac{\partial E^{\text{QM(el)/MM}}_{\text{elec}}}{\partial \mathbf{P}}, \tag{6.6}$$

where $\mathbf{V}^{\text{MM}}_{\text{m-pole}}$ is the representation of the 3-space ESP (given by Eq. (6.4)) in the AO basis $\{\omega_\mu(\mathbf{r})\}$:

$$(V^{\text{MM}}_{\text{m-pole}})_{\mu\nu} = \int \text{d}\mathbf{r}\; V^{\text{MM}}_{\text{m-pole}}(\mathbf{r}) \omega^*_\mu(\mathbf{r}) \omega_\nu(\mathbf{r}). \tag{6.7}$$

The last term in Eq. (6.1) denotes the polarization energy of the MM subsystem. (Note: the polarization of the QM system is incorporated by the SCF minimization procedure implicitly.) In the AMOEBA force field, the polarization effect is described using point induced dipoles distributed onto each MM site. With an isotropic (scalar) polarizability α_i, the induced dipole on MM site i can be expressed as

$$\begin{aligned} \boldsymbol{\mu}^{\text{ind}}_i &= \alpha_i (\boldsymbol{\mathcal{E}}^{\text{perm}}_i + \boldsymbol{\mathcal{E}}^{\text{ind}}_i) \\ &= \alpha_i (\boldsymbol{\mathcal{E}}^{\text{perm}}_i + \sum_{j \neq i}^{n_{\text{MM}}} \tilde{\mathbf{T}}^{\text{d-d}}_{ij} \boldsymbol{\mu}_j) \end{aligned} \tag{6.8}$$

$\boldsymbol{\mathcal{E}}^{\text{perm}}_i$ is the so-called "permanent" electric field. For the QM/AMOEBA system, it includes the contributions from QM nuclei and electrons, and permanent multipoles on other MM sites:

$$\boldsymbol{\mathcal{E}}^{\text{perm}}_i = \boldsymbol{\mathcal{E}}^{\text{QM(nuc)}}_i + \boldsymbol{\mathcal{E}}^{\text{QM(el)}}_i + \boldsymbol{\mathcal{E}}^{\text{MM(m-pole)}}_i. \tag{6.9}$$

Note that in the case of AMOEBA water, which is the focus of our present work, the interactions between permanent multipoles on the same fragment are *excluded*. $\boldsymbol{\mathcal{E}}^{\text{ind}}_i$, on the other hand, refers to the electric field generated by induced dipoles on every other MM site ($\tilde{\mathbf{T}}^{\text{d-d}}_{ij}$ is the Thole-damped electrostatic tensor for dipole-dipole interactions), which

requires Eq. (6.8) to be solved self-consistently. Once the induced dipoles are equilibrated, the polarization energy of the MM subsystem can be computed as

$$E_{\text{pol}}^{\text{MM}} = -\sum_{i=1}^{n_{\text{MM}}} \boldsymbol{\mu}_i^{\text{ind}} \cdot \boldsymbol{\mathcal{E}}_i^{\text{perm}} + \frac{1}{2}\sum_{ij}^{n_{\text{MM}}} \boldsymbol{\mu}_i^{\text{ind}} \cdot \mathbf{T}_{ij} \cdot \boldsymbol{\mu}_j^{\text{ind}} \tag{6.10a}$$

$$= -\frac{1}{2}\sum_{i=1}^{n_{\text{MM}}} \boldsymbol{\mu}_i^{\text{ind}} \cdot \boldsymbol{\mathcal{E}}_i^{\text{perm}} \tag{6.10b}$$

where $\mathbf{T}_{ij} = \alpha_i^{-1}\boldsymbol{\delta}_{ij} - \tilde{\mathbf{T}}_{i\neq j}^{\text{d-d}}$. The second equality holds as a result of Eq. (6.8).

In our implementation, we first variationally solve for $\{\boldsymbol{\mu}_i^{\text{ind}}(n)\}$ given the current electron density matrix $\mathbf{P}^{(n)}$ before taking the n+1 SCF step, which guarantees the stationary condition $\partial E_{\text{pol}}^{\text{MM}}/\partial \boldsymbol{\mu}_i^{\text{ind}} = 0$. Note that although the MM polarization energy can be more conveniently evaluated via Eq. (6.10b), the correct stationary condition $\sum_j \mathbf{T}_{ij}\boldsymbol{\mu}_j^{\text{ind}} = \boldsymbol{\mathcal{E}}_i^{\text{perm}}$ can only be retrieved from Eq. (6.10a) by differentiating it with respect to $\{\boldsymbol{\mu}_i^{\text{ind}}\}$. Taking advantage of this stationary condition, the contribution from MM polarization to the Fock matrix can be evaluated by differentiating Eq. (6.10a) with respect to $\mathbf{P}$:

$$\mathbf{F}_{\text{pol}}^{\text{MM}} = \left.\frac{\partial E_{\text{pol}}^{\text{MM}}}{\partial \mathbf{P}}\right|_{\{\boldsymbol{\mu}_i^{\text{ind}}\}}$$

$$= -\sum_{i=1}^{n_{\text{MM}}} \boldsymbol{\mu}_i^{\text{ind}} \cdot \frac{\partial \boldsymbol{\mathcal{E}}_i^{\text{perm}}}{\partial \mathbf{P}} \tag{6.11}$$

Among the three components on the RHS of Eq. (6.9), $\boldsymbol{\mathcal{E}}_i^{\text{QM(el)}}$ is the only density-dependent component. At MM site i, we have

$$\boldsymbol{\mathcal{E}}_i^{\text{QM(el)}} = -\nabla_{\mathbf{R}_i}\int \mathrm{d}\mathbf{r}\ \frac{\rho_{\text{el}}(\mathbf{r})}{|\mathbf{r}-\mathbf{R}_i|}$$

$$= -\operatorname{Tr}[\mathbf{P}\mathbf{V}^{\mathbf{R}_i}] \tag{6.12}$$

where $\mathbf{V}^{\mathbf{R}_i}$ is defined by

$$(V^{\mathbf{R}_i})_{\mu\nu} = \int \mathrm{d}\mathbf{r}\ \omega_\mu^*(\mathbf{r})\nabla_{\mathbf{R}_i}\frac{1}{|\mathbf{r}-\mathbf{R}_i|}\omega_\nu(\mathbf{r}). \tag{6.13}$$

Combining Eqs. (6.11) and (6.12), we have

$$\mathbf{F}_{\text{pol}}^{\text{MM}} = \sum_{i=1}^{n_{\text{MM}}} \boldsymbol{\mu}_i^{\text{ind}} \cdot \mathbf{V}^{\mathbf{R_i}}, \tag{6.14}$$

i.e., this contribution to the Fock matrix is simply the ESP generated by the current set of induced dipoles located on all MM sites.

Taking Eqs. (6.2), (6.6), and (6.14) together, the Fock matrix for the coupled QM/AMOEBA system can be expressed as

$$\mathbf{F}^{\mathrm{QM/MM}} = \mathbf{F}_0 + \mathbf{V}^{\mathrm{MM}}_{\text{m-pole}} + \sum_{i=1}^{n_{\mathrm{MM}}} \boldsymbol{\mu}_i^{\mathrm{ind}} \cdot \mathbf{V}^{\mathbf{R_i}}. \tag{6.15}$$

If we add the latter two terms in Eq. (6.15) to the core Hamiltonian ($\mathbf{H}_{\mathrm{core}}$), the one-electron energy of the QM electrons in the SCF calculation can be conveniently evaluated using $\mathrm{Tr}[\mathbf{PH}_{\mathrm{core}}]$. However, according to Eq. (6.10b), this will overcount the MM polarization energy by $-1/2\sum_i \boldsymbol{\mu}_i^{\mathrm{ind}} \cdot \boldsymbol{\mathcal{E}}_i^{\mathrm{QM(el)}}$, which thus needs to be subtracted out in the end. Therefore, the density-dependent part of the total energy for the QM/AMOEBA system (the last three terms in Eq. (6.1)) can be computed by

$$\begin{aligned} & E_{\mathrm{el}}^{\mathrm{QM}} + E_{\mathrm{elec}}^{\mathrm{QM(el)/MM}} + E_{\mathrm{pol}}^{\mathrm{MM}} \\ &= \mathrm{Tr}[\mathbf{PH}_{\mathrm{core}}] + \frac{1}{2}\mathrm{Tr}[\mathbf{PIIP}] + E_{\mathrm{xc}} + \frac{1}{2}\sum_{i=1}^{n_{\mathrm{MM}}} \boldsymbol{\mu}_i^{\mathrm{ind}} \cdot \boldsymbol{\mathcal{E}}_i^{\mathrm{QM(el)}} \\ &- \frac{1}{2}\sum_{i=1}^{n_{\mathrm{MM}}} \boldsymbol{\mu}_i^{\mathrm{ind}} \cdot (\boldsymbol{\mathcal{E}}_i^{\mathrm{QM(nuc)}} + \boldsymbol{\mathcal{E}}_i^{\mathrm{MM(m\text{-}pole)}}), \end{aligned} \tag{6.16}$$

where **II** represents the two-electron AO integrals that are used for the construction of Coulomb and exact exchange (if hybrid functionals are employed) matrices, and E_{xc} is the exchange-correlation (XC) energy of KS-DFT. We note that the SCF equations presented above for mutually polarizable QM/AMOEBA calculations are consistent with those reported by Loco et al.[478] in a previous paper and also similar to many other QM/MMpol studies (e.g., the early work by Thompson and Schenter [492]).

With equations for the total energy and the corresponding Fock matrix available, we adopt a double-loop SCF optimization scheme to minimize the total energy of the QM/AMOEBA system, i.e., E_{tot} is optimized with respect to both QM electron density (outer loop) and AMOEBA induced dipoles (inner loop). The entire procedure of this double-loop SCF calculation is as follows:

1. Compute the contribution from AMOEBA permanent multipoles to the core Hamiltonian (Eq. (6.7)), and the electric field matrix $\mathbf{V}^{\mathbf{R}_i}$ for each inducible MM site i (Eq. (6.13)).

2. Obtain the initial guess for the QM electron density matrix ($\mathbf{P}$).

3. With the given $\mathbf{P}$, iteratively solve for the induced dipoles in the MM region ($\{\boldsymbol{\mu}_i\}$) within the inner loop.

4. Evaluate the contribution from induced dipoles to the core Hamiltonian (Eq. (6.14)).

5. Build the Fock matrix for the current (outer loop) iteration according to Eq. (6.15), and evaluate the SCF energy (Eq. (6.16)) and the error vector or gradient.

6. Check for convergence:

   - If NOT converged, update the MO coefficients and the density matrix with the employed SCF algorithm, and go back to Step 3.
   - If converged, compute the other energy components in Eq. (6.1) that are density-independent (e.g. the vdW interaction between QM and MM regions).

6.2.2 Energy decomposition analysis

As the first assessment of this mutually polarizable and fully self-consistent QM/AMOEBA model, we apply it to the evaluation of solute-solvent interactions (the solute is described by QM and the solvent molecules by AMOEBA) and compare the results to full QM reference calculations. In our previous study (using the ONETEP/TINKER implementation of QM/AMOEBA), [168] it has been revealed that the unmodified QM/AMOEBA model underestimates solute-solvent interaction energies almost consistently across a range of systems, and further softening the buffered 14-7 potential for $E_{\text{vdw}}^{\text{QM/MM}}$ significantly improves the agreement with the full QM reference for most of the investigated systems. Nonetheless, the primary reason for the poor performance of the unmodified QM/AMOEBA model is not completely clear without further analysis. Here, similar to a recent benchmark of ion-water interactions that we performed for the AMOEBA force field (pure MM) against the ALMO-EDA results,[167] we decompose the solute-solvent interactions evaluated by both QM/AMOEBA and full QM into contributions from permanent electrostatics (ELEC), polarization (POL), and vdW interaction (vdW), and the agreement in total interaction energy and in each energy component will be assessed.

Since our QM/AMOEBA model contains mutual polarization, the interaction energy between QM and MM regions ($E_{\text{int}}^{\text{QM/MM}}$) cannot be simply represented by terms in Eq. (6.1). Instead, it can always be computed through the so-called supermolecular approach, i.e.,

$$E_{\text{int}}^{\text{QM/MM}} = E_{\text{tot}}^{\text{QM/MM}} - E_0^{\text{QM}} - E_0^{\text{MM}}, \tag{6.17}$$

where E_0^{QM} and E_0^{MM} refer to the total energies of isolated QM and MM subsystems, respectively.

With the goal of extracting the energy components of this interaction across the QM/MM boundary, we propose the following approach which has a similar spirit to that of the ALMO-EDA. First, we perform SCF calculations for isolated QM and MM subsystems, and the resulting QM density matrix ($\mathbf{P}_0$) and MM induced dipoles ($\{\boldsymbol{\mu}_i^{\text{ind}}(0)\}$) are collected and employed as the initial guess for the coupled QM/AMOEBA calculation. Note that the MM induced dipoles and $E_{\text{pol}}^{\text{MM}}$ are both non-zero at this stage unless there is only one single polarization group (equivalent to one single fragment for the AMOEBA water case) in the MM region, whereas the polarization effect arises from the existence of MM multipoles *only*. Then the QM and MM subsystems are coupled together, and before any further relaxation occurs, the permanent electrostatic interaction between QM and AMOEBA fragments can

be computed as

$$E_{\text{elec}}^{\text{QM/MM}} = \text{Tr}[\mathbf{P}_0 \tilde{\mathbf{V}}_{\text{m-pole}}^{\text{MM}}] + \tilde{E}_{\text{elec}}^{\text{QM(nuc)/MM}}, \tag{6.18}$$

where the tildes indicate that interaction between the initial set of induced dipoles ($\{\boldsymbol{\mu}_i^{\text{ind}}(0)\}$) and the QM region (electron density and nuclei) is also included. At this point, the induced dipoles are computed *without* coupling with the QM subsystem, so they play the same role as permanent dipoles and can be treated in the same way computationally, i.e., $\tilde{\mathbf{V}}_{\text{m-pole}}^{\text{MM}}$ can be computed based on Eqs. (6.4) and (6.7) by simply replacing $\boldsymbol{\mu}_i$ with $\boldsymbol{\mu}_i' = \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^{\text{ind}}(0)$. This definition of permanent electrostatics in mutually polarizable QM/MM calculations, especially the use of $\{\boldsymbol{\mu}_i^{\text{ind}}(0)\}$, is the same as the scheme proposed in Ref. 509.

We then relax the QM/AMOEBA system following the SCF procedure presented above, and the resulting changes in the three terms on the RHS of Eq. (6.16) define the polarization energy of the whole QM/MM system:

$$E_{\text{pol}}^{\text{QM/MM}} = \Delta E_{\text{el}}^{\text{QM}} + \Delta E_{\text{elec}}^{\text{QM(el)/MM}} + \Delta E_{\text{pol}}^{\text{MM}}. \tag{6.19}$$

Note that due to the mutual character of polarization here, it is *not* possible to further decompose $E_{\text{pol}}^{\text{QM/MM}}$ into forward and backward contributions while retaining additivity. Once the SCF is converged, the vdW interaction between QM and MM subsystems, $E_{\text{vdw}}^{\text{QM/MM}}$, can be evaluated at the end since it is fully classical in our current model. Taken together, the interaction across the QM/MM boundary is decomposed into three energy components:

$$E_{\text{int}}^{\text{QM/MM}} = E_{\text{elec}}^{\text{QM/MM}} + E_{\text{pol}}^{\text{QM/MM}} + E_{\text{vdw}}^{\text{QM/MM}}. \tag{6.20}$$

For the purpose of the benchmark, solute-solvent interaction energies evaluated by full QM also need to be decomposed into the corresponding energy components. This can be achieved by applying a slightly modified version of the ALMO-EDA, which has been introduced in Ref. 167. To briefly recapitulate the modifications, we adopt the "classical" definition for permanent electrostatics, i.e., the classical Coulomb interaction between charge distributions of fragments evaluated in isolation:

$$E_{\text{elec}} = \sum_{A<B} \int_{\mathbf{r}_1} \int_{\mathbf{r}_2} \rho_A^{\text{tot}}(\mathbf{r}_1) \frac{1}{r_{12}} \rho_B^{\text{tot}}(\mathbf{r}_2) \text{d}\mathbf{r}_1 \text{d}\mathbf{r}_2, \tag{6.21}$$

where $\rho_A^{\text{tot}}(\mathbf{r}) = \rho_A^{\text{el}}(\mathbf{r}) + \rho_A^{\text{nuc}}(\mathbf{r})$. Here ρ_A and ρ_B simply refer to charge distributions of solute and solvent molecules, respectively. Then, with the definition for the dispersion energy (E_{disp}) remaining unchanged (based on Ref. 118), the remainder of the frozen interaction energy (interaction between fragments approaching each other whose molecular orbitals are *not yet* relaxed) is defined as the (modified) Pauli term:

$$E_{\text{pauli}} = E_{\text{frz}} - E_{\text{elec}} - E_{\text{disp}}. \tag{6.22}$$

The separation of polarization and charge transfer (CT) still utilizes the so-called "SCF for molecular interaction" (SCF-MI) approach. Basis set superposition error (BSSE) evaluated

by the counterpoise correction can also be included if desired, which is often combined with the CT term because they both arise from the delocalization effect. Thus the solute-solvent interaction energy evaluated by full QM can be decomposed as follows:

$$E_{\text{int}} = E_{\text{elec}} + E_{\text{pauli}} + E_{\text{disp}} + E_{\text{pol}} + E_{\text{ct}}. \tag{6.23}$$

The readers are referred to the original references of ALMO-EDA (Refs. [97, 115, 118]) for more details about this method.

In order to further simplify the comparison between energy components evaluated by QM/AMOEBA and full QM, in this work we introduce another "coarse-grained" modification to the ALMO-EDA scheme presented above, i.e., we define the "vdW" component of the full QM interaction energy as the sum of Pauli repulsion, dispersion and charge transfer:

$$E_{\text{vdw}} = E_{\text{pauli}} + E_{\text{disp}} + E_{\text{ct}}, \tag{6.24}$$

and then Eq. (6.23) becomes

$$E_{\text{int}} = E_{\text{elec}} + E_{\text{pol}} + E_{\text{vdw}}, \tag{6.25}$$

which has a straightforward correspondence to the decomposition of QM/AMOEBA interaction energy represented by Eq. (6.20). This choice, especially the addition of CT to the other two energy components that correspond to the vdW interaction physically, is rationalized by our previous investigation of vdW interactions in AMOEBA with the same EDA scheme. [167]

6.2.3 Implementation through the Q-Chem/LibEFP interface

The QM/AMOEBA model presented above and the energy decomposition analysis scheme for interactions across the QM/MM boundary are implemented through a Q-Chem/LibEFP code structure. LibEFP [507, 508] is an open source library for effective fragment potential (EFP) [153, 154, 511] calculations and has been interfaced with the Q-Chem 4 software package.[255] It can be naturally extended to support the AMOEBA force field thanks to the similarities between these two models, e.g., they both use distributed point multipoles to model permanent electrostatics and use distributed inducible dipoles to describe polarization. Therefore, many routines implemented for EFP calculations, such as those that are utilized to compute the interactions between point multipoles, can be directly used for AMOEBA. Nonetheless, several additional functionalities are still required in order to fully support AMOEBA in LibEFP, including

- Parsers for AMOEBA parameter files which contain permanent multipole moments, atomic polarizabilities, vdW parameters, etc.;
- The valence terms in the AMOEBA force field (bond, angle and Urey-Bradley), which were not included in LibEFP since fragments with fixed geometries are usually used in EFP calculations;

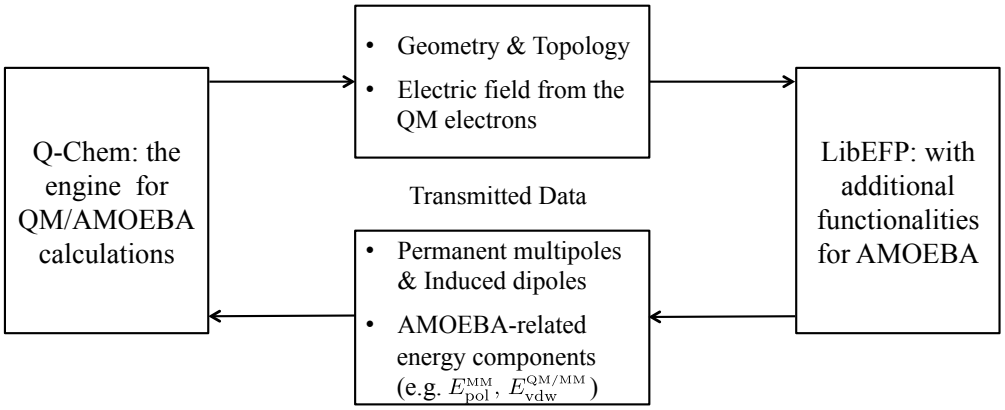

Figure 6.1: The Q-Chem/LibEFP code interface for mutually polarizable, fully self-consistent QM/AMOEBA calculations.

- vdW interactions between AMOEBA fragments and between QM and MM atoms, which are both described by the buffered 14-7 potential;
- Routines transforming AMOEBA's permanent multipoles that are in their own local coordinates (as in the parameter file) into the global coordinate, which is a necessary step before including these permanent multipoles into the QM/MM system;
- Thole-damped electric field arising from monopole, dipole, and quadrupole moments, which is required by the evaluation of induced dipoles based on Eq. (6.8).

Calculations using the AMOEBA force field can be correctly handled when these additions are combined with the original routines in LibEFP. We note that currently we only enabled the use of the AMOEBA water model in LibEFP, although a future extension to more general MM systems described by AMOEBA should be straightforward.

The Q-Chem/LibEFP code interface for mutually polarizable QM/AMOEBA calculations is illustrated in Figure 6.1. Q-Chem serves as the driver for the entire calculation, which parses the geometry and topology (connectivity between the MM atoms) and passes the information to LibEFP. LibEFP computes MM energy terms (such as $E_{\rm val}^{\rm MM}$, $E_{\rm elec}^{\rm MM}$) and components of QM/MM interaction energy that are not density-dependent (such as $E_{\rm vdw}^{\rm QM/MM}$ and $E_{\rm elec}^{\rm QM(nuc)/MM}$), and it also passes AMOEBA's permanent multipoles (all transformed into the global coordinate frame) and current set of induced dipoles to Q-Chem. Since routines that evaluate electric potential, field, and field derivative matrices in the AO basis are already available in Q-Chem, only slight modifications to the standard SCF routines are needed to incorporate the MM contributions to the Fock matrix. And standard SCF algorithms, such as the direct inversion of the iterative subspace (DIIS) method,[8, 9] can still be employed for minimizing the energy of a coupled DFT/AMOEBA system. In terms of the MM polarization energy ($E_{\rm pol}^{\rm MM}$), the induced dipoles of AMOEBA are self-consistently solved in LibEFP,

while the electric field component arising from QM electron density on each inducible site is evaluated by and then passed from Q-Chem.

6.3 Results

6.3.1 Computational details

All the QM/AMOEBA and full QM calculations are performed with a locally modified Q-Chem 4.4 software package, [255] which has been interfaced with a locally developed version of LibEFP. [507, 508] In this work, "QM" refers to DFT methods *exclusively*, and unless otherwise specified, the ωB97X-V functional [72] is used. ωB97X-V is a range-separated hybrid GGA which employs the VV10 non-local correlation (NLC) functional [71] to describe dispersion, and its accuracy for non-covalent interactions, especially for those involving ionic species, has been shown by several recent studies. [75, 76, 167] Two Karlsruhe basis sets, def2-SVPD and def2-TZVPPD, [183] are employed in the calculations presented below, which are of augmented double- and triple-ζ quality, respectively.

The full QM and QM/AMOEBA solute-solvent interaction energies are both evaluated with the supermolecular approach, and counterpoise corrections for BSSE are applied to the former. The ALMO-EDA calculations are performed based on the modified scheme introduced in Sec. 6.2.2, while in this work, the separation of POL and CT is achieved by using the original AO-block based ALMO scheme [115, 234] instead of the recently proposed fragment electrical response function (FERF) model.[117] The FERF model gives a well-defined basis set limit to the resulting polarization energy but is computationally more expensive. This choice is made because smaller basis sets that are far from the complete basis set (CBS) limit are used in this work, and it is known that these two models will *not* yield remarkably different results unless very large basis sets (usually beyond augmented triple-ζ) are used. [117] As a validation, we compare the polarization energies obtained by using the ALMO and FERF models with the def2-TZVPPD basis for several systems that are investigated below (the results are shown in Sec. E.1), and the differences between them are shown to be insignificant.

The AMOEBA 03 water model [139] is utilized for the solvent molecules in all QM/MM calculations, while the vdW parameters for the solute atoms are taken from the most recently released parameter files of AMOEBA ("amoebapro13" or "amoeba09") in the TINKER 7 molecular modeling package.[413] The induced dipoles are converged to 10^{-10} a.u. in each inner loop, while the convergence criterion for the outer loop SCF is 10^{-8} a.u. We note that for each inducible site, while the polarizing effect of permanent multipoles and other induced dipoles are evaluated with Thole damping, the electric field arising from QM nuclei and electrons are computed *without* applying any damping schemes. Furthermore, the polarization of QM electron density due to the embedding potential of MM is also evaluated *without* any damping. Therefore, the QM/AMOEBA model employed here stands for a preliminary, unmodified coupling between the QM and MM components, and the necessity

of applying damping functions for forward or backward polarizations will be examined via EDA calculations.

6.3.2 The water dimer and the water-Cl^- complex

We start by investigating the performance of QM/AMOEBA on the prototypical system for hydrogen-bonding interaction — the water dimer, where each water molecule can be treated with either QM or AMOEBA. Its dissociation potential energy surfaces (PES) evaluated with full QM, full AMOEBA, and QM/AMOEBA are shown in the top left panel of Figure 6.2, as functions of the O··O distance between two water molecules. The equilibrium geometry is optimized at the ωB97X-V/def2-QZVPPD level of theory, and the PES scan is performed by modifying the O··O distance only without relaxing other coordinates (the PES scan for the water-Cl^- system below is performed in the same manner). The first thing to note is that the choice of basis set makes a minimal difference to the interaction energies (and their components) computed by both full QM and QM/AMOEBA, which is shown in Figure E.3 (in Appendix E.2). Therefore, we focus on the results evaluated with the def2-TZVPPD basis in the following discussion of this system. In the scenario where the H-donor water molecule (water 1) is described by QM and the H-acceptor (water 2) by AMOEBA, the PES yielded by QM/AMOEBA (green curve) is slightly underbound at all ranges measured against the full QM reference, and the difference at the common minimum-energy distance (2.90 Å) is 0.3 kcal/mol. Around the minimum, the full AMOEBA PES (red dashed) shows better agreement with the full QM reference, while in the short range where pure AMOEBA considerably underbinds the water dimer, QM/AMOEBA deviates from the full QM PES less.

The character of the QM/AMOEBA PES can be elucidated with the EDA results. According to the top right panel of Figure 6.2, the permanent electrostatic interaction between water 1 (QM) and 2 (AMOEBA) is marginally less attractive compared to that in the full AMOEBA case. As we have shown in Ref. 167, the charge penetration error (CPE), which can be interpreted as the difference between AMOEBA's multipole-based permanent electrostatics and the full QM counterpart, is significant for the water dimer at the equilibrium distance, and the good agreement in the total energy around equilibrium relies on the cancellation of errors through the softened vdW potential in AMOEBA, which is shown in the bottom right panel of Figure 6.2. This argument applies to the QM/AMOEBA case as well since it has a similar profile for permanent electrostatics, and the special characteristics of QM/AMOEBA's total PES can be attributed to the polarization component. At the ranges around and beyond equilibrium, the polarization energies given by pure AMOEBA and ALMO-EDA agree almost perfectly, whereas that given by QM/AMOEBA is slightly underestimated, which appears to be the primary reason for the underbinding of QM/AMOEBA in this region. However, in the short range the polarization energy of full AMOEBA becomes unphysical due to the onset of Thole damping and that results in the overly repulsive total interaction energy, while the profile given by QM/AMOEBA shows a more proper behavior

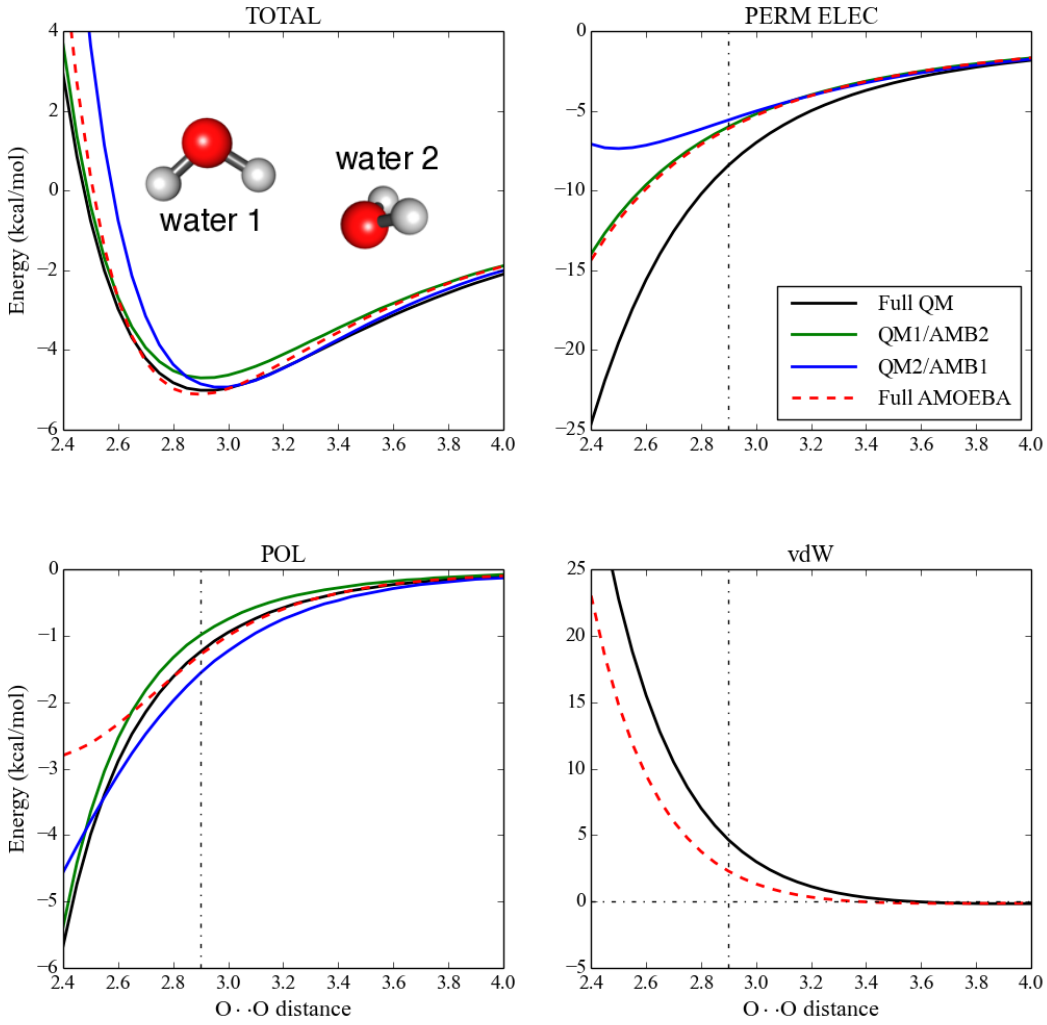

Figure 6.2: Total interaction energies and their components (in kcal/mol) evaluated with full QM, full AMOEBA, and QM/AMOEBA for the water dimer as functions of the O··O distance (the interval between two neighboring data points is 0.05 Å). The def2-TZVPPD basis set is employed for the QM and QM/AMOEBA calculations. Water 1 and 2 refer to the proton donor and acceptor, respectively, whose geometries are unrelaxed during the PES scan. "QM1/AMB2" refers to the scenario where water 1 is described by QM and water 2 by AMOEBA, and "QM2/AMB1" is defined in the same way. The vdW interactions in QM/AMOEBA calculations are identical to those computed by pure AMOEBA so they are not plotted in the bottom right panel.

despite still being insufficiently favorable, which contributes to its better agreement with full QM in this region in terms of total interaction energy.

It is intriguing to examine the performance of QM/AMOEBA after switching the QM and MM regions, i.e., we describe the H-acceptor (water 2) with QM and the H-donor (water 1) with AMOEBA instead. This, referred to as the "reversed" QM/AMOEBA model for the water dimer in the discussions below, yields a rather skewed total potential energy profile (blue curve in Figure 6.2), where the equilibrium O··O distance is shifted to 2.95–3 Å and the repulsive wall becomes even harder than the pure AMOEBA case. Turning to the energy components, the first striking result is that the permanent electrostatics deviates from the values given by pure AMOEBA and the former QM/AMOEBA model enormously after entering the compressed region (r(O··O) $<$ 2.9 Å). At r(O··O) = 2.5 Å, it differs from pure AMOEBA's permanent electrostatic interaction by over 4 kcal/mol, and below that distance it becomes less attractive with decreasing intermolecular distances, which is physically incorrect and further enlarges the gap. Recalling that the vdW potential employed here is parameterized with the CPE in the pure AMOEBA case taken into account, it will thus be insufficient to address an even larger CPE here, and this is the primary reason for the excessively hard repulsive wall given by the "reversed" QM/AMOEBA model. Its polarization energy, contrary to that given by the former QM/AMOEBA model, is more favorable than the full QM reference except at the most compressed distances (r(O··O) $\leq$ 2.5 Å). Nevertheless, the difference in polarization energy has a small magnitude so that it barely affects the error dominated by CPE.

In order to see if this represents the typical situation when AMOEBA water serves as the proton donor in H-bonding systems, we also investigate the performance of QM/AMOEBA for the rigid dissociation PES of the water-Cl^- complex (the chloride anion is treated with QM), which is closely related to the "reversed" water dimer case discussed above. Its total energy and EDA results are shown in Figure 6.3. Note that although the choice of basis set has a slightly larger effect on this system (which can be seen from the "POL" panel of Figure E.4 in Appendix E.2), it brings no qualitative changes so that we can still focus on the def2-TZVPPD results. Indeed, when the (QM) H-acceptor water is replaced by Cl^- which interacts with the AMOEBA water more strongly, the issues revealed by the investigation of the water dimer above are further exacerbated, which gives rise to an enormously shifted PES. In contrast to the pure AMOEBA case, QM/AMOEBA underbinds the H_2O··Cl^- complex by 3.6 kcal/mol at the QM minimum (3.10 Å), and the error further increases in the compressed region. Meanwhile, the equilibrium Cl··O distance given by QM/AMOEBA (3.35 Å) is also substantially elongated compared to the full QM result.

Turning to the energy components, the permanent electrostatics of QM/AMOEBA shows a similar but more pronounced feature as in the "reversed" water dimer case, as it starts to deviate from the pure AMOEBA curve considerably at an even longer distance, and the difference at the QM equilibrium distance is already as large as 7 kcal/mol. This is evidently the main culprit for the significant underbinding of QM/AMOEBA at all ranges. Although the polarization energy and vdW interaction given by QM/AMOEBA are both more favorable than their full QM counterparts, they are far from being sufficient to cancel

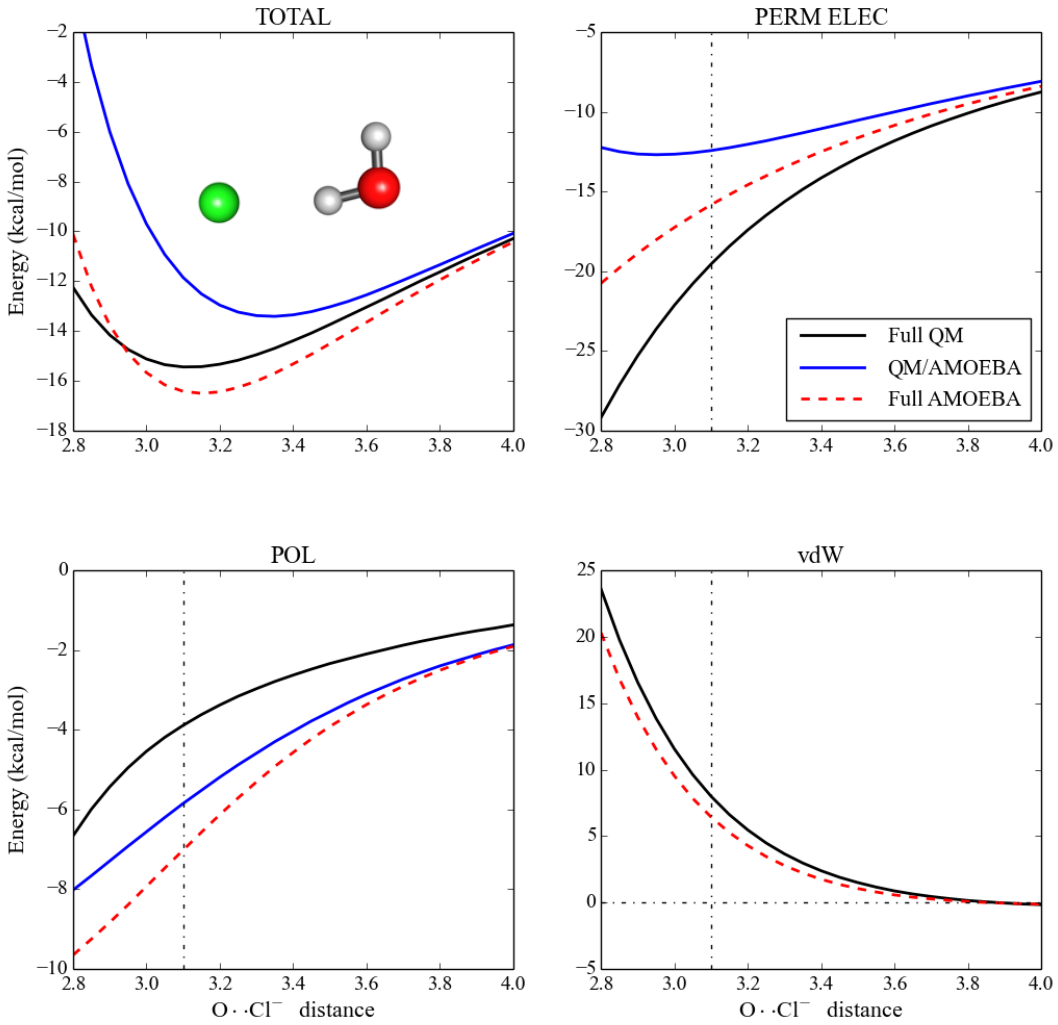

Figure 6.3: Total interaction energies and their components (in kcal/mol) for the water-Cl^- complex as functions of the O··Cl^- distance. In QM/AMOEBA calculations, Cl^- is treated with QM and the H_2O molecule is described by AMOEBA. The other computational and plotting details are the same as in Figure 6.2.

the enormous error in permanent electrostatics. It should be noted that the undamped QM/AMOEBA polarization energy lies in between the full QM and full AMOEBA results across the board, i.e., it still overestimates the polarization energy relative to the full QM reference but less severely than pure AMOEBA, even though Thole damping is consistently applied to the latter.

Taking these two examples (the "reversed" water dimer and the water-Cl^- complex) together, it is clearly revealed that QM/AMOEBA can suffer more from CPE than pure AMOEBA, and that the resulting profile of permanent electrostatics, instead of lying in between the full QM and full AMOEBA ones, can be *less attractive than both*. This is slightly counterintuitive and will be further discussed in Sec. 6.3.5.

6.3.3 Interaction with solvent molecules in the first solvation shell

As the first step from gas-phase dimers to clusters in condensed phase, we turn to the interaction between solutes and solvent (water) molecules in their first solvation shells. Three solutes are considered in this study: H_2O, Na^+, and Cl^-, which are representative of neutral, cationic, and anionic species, respectively. The configurations are taken from MD simulations of one solute molecule solvated in a box of 215 H_2O molecules using the AMOEBA force field, and the details regarding the equilibration and production steps of the MD simulation are the same as in the previous work by Grossfield et al. [385] For each species, the number of water molecules in the first solvation shell (coordination number) is determined by an integration over the first peak of the resulting radial distribution function (RDF) until the position of the first minimum, and the average (closest integer value) turns out to be 4 for H_2O, 6 for Na^+, and 8 for Cl^-. The resulting coordination numbers for Na^+ and Cl^- are consistent with the values reported in Ref. 512 (note: the coordination numbers reported in Ref. 385 were not calculated from the RDFs), while they might differ from those obtained from AIMD simulations performed by others.[454, 459, 460] Nonetheless, as our focus here is to benchmark solute-solvent interaction energies rather than the solvent structure, the configurations prepared by classical MD should be able to fulfill this purpose. To simplify the discussions below, for each solute species, the same number of nearest water molecules (the average integer coordination numbers given above) is applied to all its configurations. Therefore, these clusters are also referred to as $(H_2O)_5$ (water pentamer), $Na^+(H_2O)_6$ and $Cl^-(H_2O)_8$ in the discussions below.

The solute-solvent interaction energies evaluated with QM/AMOEBA are plotted with respect to the full QM results in the top row of Figure 6.4. For the H_2O and Na^+ solutes, the QM/AMOEBA interaction energies agree with the full QM results reasonably but are slightly underbound, where the mean signed errors (MSEs) are +1.29 kcal/mol and +1.44 kcal/mol, respectively (the statistical errors for each solute are shown in Table 6.1). Note that the "solute" water molecule can play the role of H-donor and H-acceptor simultaneously, and according to the results above for the water dimer (Figure 6.2), QM/AMOEBA underes-

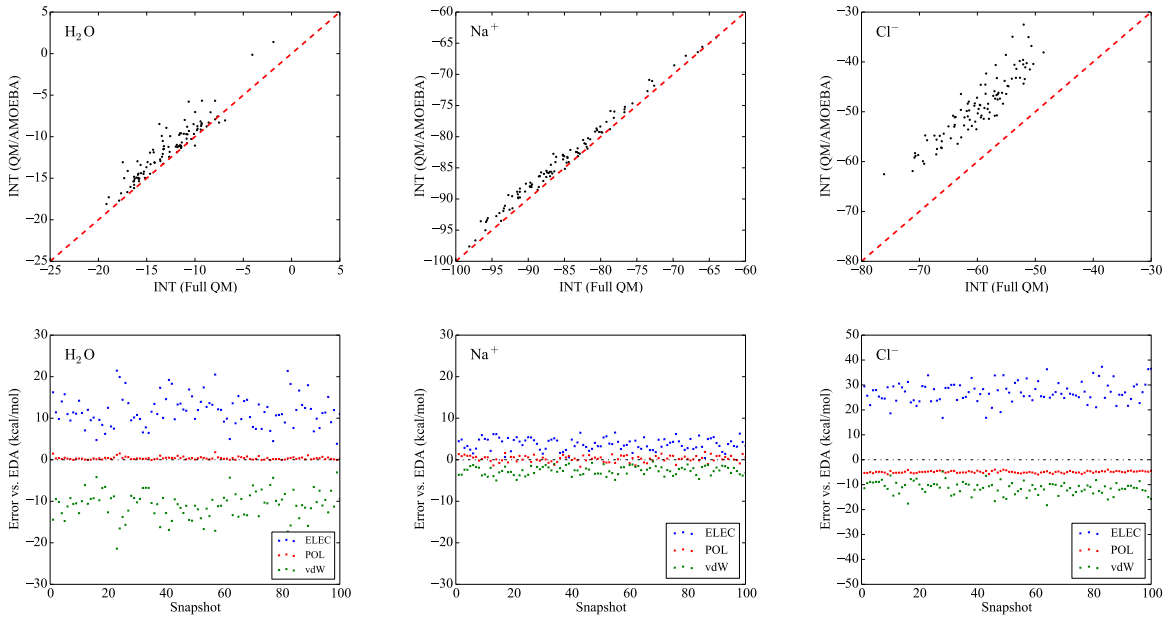

Figure 6.4: First row: interaction energies (in kcal/mol) between three solutes (H_2O, Na^+, Cl^-) and water molecules in their first solvation shells evaluated with full QM (x) and QM/AMOEBA (y); second row: errors (in kcal/mol) for the three components of QM/AMOEBA interaction energies (ELEC, POL, vdW) measured against the reference values given by ALMO-EDA. 100 snapshots are calculated and then plotted for each solute-solvent system.

timates the binding energy near equilibrium in either case, which seems to be consistent with the general trend here where a QM water molecule is interacting with four AMOEBA waters in the vicinity. On the other hand, the agreement between QM/AMOEBA and full QM for Cl^- is poor, where we see a substantial underbinding by QM/AMOEBA (the MSE over 100 snapshots is +10.94 kcal/mol). This result is also consistent with what we have found for the dissociation PES of the water-Cl^- complex, where QM/AMOEBA underestimates the interaction energy more significantly than in the water dimer case at their individual equilibrium distances.

The components of QM/AMOEBA interaction energies are compared against their full QM counterparts as well, and the errors are shown in the bottom row of Figure 6.4. Distinct patterns emerge for each of these species. For the cationic solute (Na^+), the mutual polarization of QM/AMOEBA (which is dominated by solvent polarization in this case) matches that given by ALMO-EDA closely, where the mean absolute error (MAE) is only 0.65 kcal/mol. The errors in permanent electrostatics and vdW interaction are exclusively above and below zero, respectively, which largely cancel each other but not completely, yielding the slightly underbound results compared to full QM. Turning to the neutral solute H_2O, while the small deviation from ALMO-EDA's polarization energy is retained, the errors in permanent electrostatics and vdW become considerably larger (see the statistics in Table 6.1). The errors

Table 6.1: Maximum errors (MAX), mean absolute errors (MAE), and mean signed errors (MSE) of QM/AMOEBA (in kcal/mol) for interactions between three solutes (H_2O, Na^+, and Cl^-) and the H_2O molecules in their first solvation shells. The statistical errors in total energies and energy components are evaluated relative to the full QM references over 100 samples (snapshots) for each solute.

		MAX	MAE	MSE
H_2O	INT	5.21	1.35	1.29
	ELEC	21.46	11.79	11.79
	POL	1.79	0.37	0.36
	vdW	-21.40	10.86	-10.86
Na^+	INT	3.78	1.45	1.44
	ELEC	6.55	3.79	3.79
	POL	1.89	0.65	0.25
	vdW	-4.99	2.60	-2.60
Cl^-	INT	19.51	10.94	10.94
	ELEC	37.32	27.17	27.17
	POL	-5.94	4.91	-4.91
	vdW	-18.25	11.33	-11.33

also span much larger ranges than in the former Na^+ case, which can be seen from the contrast between the first two bottom panels in Figure 6.4. Nevertheless, thanks to the rather similar magnitude of the positive and negative errors, the resulting total interaction energies are brought to a reasonable level. This, however, does not hold for the anionic solute Cl^-, and the features demonstrated by EDA are in line with our expectations based on the results for the water-Cl^- complex. The major source of error is clearly the permanent electrostatics, which on average is less attractive than its full QM counterpart by 27.17 kcal/mol, and the vdW potential of AMOEBA is evidently not sufficiently softened to compensate for such a large error. Also, as opposed to the other two solutes, the polarization energies given by QM/AMOEBA for $Cl^-(H_2O)_8$ clusters are systematically overestimated (on average it is more favorable by 4.9 kcal/mol than the full QM references), which is also consistent with the trend demonstrated in Figure 6.3.

The results above indicate that the performance of this QM/AMOEBA model in describing the interaction between solute and solvent molecules in its vicinity strongly depends on the balance of the QM/MM interface, i.e., the error cancellation between individual components of the interaction crossing the boundary of QM and MM regions. Since the error in polarization energy usually has the smallest magnitude, the balance between inadequately attractive electrostatic interaction and softened, fully empirical vdW potential has a decisive influence on the accuracy of solute-solvent interaction energies computed with QM/AMOEBA.

6.3.4 Two case studies: $F^-(H_2O)_{10}$ and $CH_4(H_2O)_{20}$ clusters

In order to verify the generality of our findings above, we investigate two other solute-water clusters, $F^-(H_2O)_{10}$ (including ten isomers) and $CH_4(H_2O)_{20}$ (methane in a dodecahedral cage formed by water molecules), whose geometries are taken from Ref. 276 and 513, respectively. These systems have been previously studied by Lao et al. with a broad range of quantum chemistry methods.[76] It should be noted that all ten water molecules in $F^-(H_2O)_{10}$ are treated as one single monomer here so that the binding energies reported below will not be comparable to those in Ref. 76, where each H_2O was treated as an individual monomer.

The total interaction and energy decomposition results for 10 isomers of $F^-(H_2O)_{10}$ are shown in Figure 6.5 (the original data are provided in Table E.3), and the manifested features are similar to those of the $Cl^-(H_2O)_8$ systems studied above. The plot for the total interaction energies given by QM/AMOEBA is almost parallel with that for the full QM values, and according to the original data, QM/AMOEBA reproduces the correct ordering of these isomers except that it computes the binding energy for Isomer 8 too low relative to Isomer 1 and 2. However, there exists an over 20 kcal/mol gap between the absolute binding energies given by QM/AMOEBA and full QM, due to the highly unbalanced QM/MM interface. Figure 6.5 shows the tremendous discrepancy between their permanent electrostatics, which is about 60 kcal/mol, whereas the vdW interactions given by AMOEBA's 14-7 potential are only about 30 kcal/mol below the full QM values. Also, like other H-bonding systems where AMOEBA water donates its proton to the QM region, the polarization energies for these isomers are overestimated by about 7 kcal/mol, which is slightly larger than the MAE for the $Cl^-(H_2O)_8$ clusters.

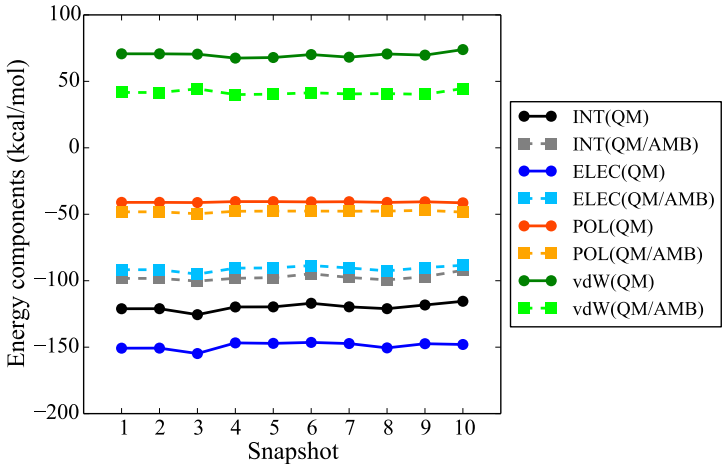

Figure 6.5: Total interaction energies and energy components (in kcal/mol) computed with full QM and QM/AMOEBA for F^- interacting with 10 surrounding water molecules in 10 isomers of the $F^-(H_2O)_{10}$ cluster.

Methane, on the other hand, is a non-polar solute so that the contributions from permanent electrostatics and polarization to the interaction of CH_4 with the dodecahedral water cage are expected to be weaker. The total interaction energy given by QM/AMOEBA is -5.09 kcal/mol, which is about 0.8 kcal/mol above the full QM result. Nevertheless, according to Table 6.2, the components of full QM and QM/AMOEBA interactions manifest sharp discrepancies. With full QM, permanent electrostatics and polarization make non-trivial contributions to binding, especially the former which accounts for roughly 50% of the stabilization, while in the QM/AMOEBA case, their contributions appear to be trivial and the whole system is almost entirely bound by QM/MM vdW interaction which is described at the AMOEBA level. We note that contrary to the former $F^-(H_2O)_{10}$ system, the AMOEBA waters are no longer proton donors here and the resulting QM/AMOEBA polarization energy is largely underestimated. It is thus evident that the electrostatic interaction (permanent and induced) between a neutral, non-polar solute and solvent water molecules cannot be correctly described by the present QM/AMOEBA model which has no explicit treatment for the charge penetration effect, and the quality of the resulting total interaction energy is entirely controlled by the empirical 14-7 potential.

Table 6.2: Total interaction energies and their components (in kcal/mol) for methane interacting with a dodecahedral water cage ($CH_4(H_2O)_{20}$) evaluated with full QM and QM/AMOEBA.

	INT	ELEC	POL	VDW
Full QM	-5.90	-2.90	-0.56	-2.44
QM/AMOEBA	-5.09	-0.08	-0.03	-4.98

6.3.5 More detailed investigations on electrostatics and polarization in H-bonding systems

According to the examples investigated above, we have noticed that the most challenging systems for QM/AMOEBA are H-bonding complexes where AMOEBA water serves as the proton donor. For these systems, one of the most notable features of QM/AMOEBA is that the resulting permanent electrostatics is even less favorable than that given by pure AMOEBA. In order to understand this better, we perform a further analysis on the electrostatic interaction between Cl^- and H_2O at the equilibrium geometry of the water-Cl^- complex, where the contributions from nuclei and electrons are separated. The results are shown in Table 6.3. Note that for atomic site i in the MM region, the "electron" part includes a point charge of value $q_i - N_i$ (q_i is the original AMOEBA monopole for site i and N_i is the corresponding nuclear charge) and all the higher-order multipoles. Such a definition ensures that the "nuclei" part in QM and MM regions are treated equivalently. According to Table 6.3, when the QM water is replaced by AMOEBA, the interactions between the electron part of H_2O with the nucleus ($+17e$) and electron density of Cl^- both become less

attractive, and the substantial deviation of QM/AMOEBA's permanent electrostatics relative to the full QM reference primarily arises from the overly unfavorable electron-electron component. Turning to the pure AMOEBA case, where the electron density of Cl^- collapses onto the same position as its nucleus and reduces to a point charge (-18e), the electron-electron part becomes even more repulsive. However, the attractive interaction between the electrons of Cl^- and the nuclei of H_2O is also enhanced upon this change, which compensates for the overly repulsive electron-electron component to a large extent and gives rise to its more favorable electrostatic interaction than that of QM/AMOEBA. Here we see that the charge penetration effect, as reflected by the lack of attractiveness in permanent electrostatics, involves the interplay of three distinct energy components.

Table 6.3: Components of permanent electrostatic interaction (in kcal/mol) for the $H_2O{\cdot\cdot}Cl^-$ complex at its equilibrium geometry, as computed with full QM, full AMOEBA, and QM/AMOEBA. "n" refers to nuclei, "e" represents electrons, and the numbers reported in parentheses are errors with regard to the full QM values. The nuclei-nuclei component is identical in all three calculations (18808.74 kcal/mol) so it is not listed in the table.

	Cl^-(n)$\cdot\cdot H_2O$(e)	Cl^-(e)$\cdot\cdot H_2O$(n)	Cl^-(e)$\cdot\cdot H_2O$(e)
Full QM	-18544.86	-19901.80	19618.83
Full AMOEBA	-18544.12 (+0.74)	-19915.14 (-13.34)	19634.95 (+16.12)
QM/AMOEBA	-18544.12 (+0.74)	-19901.80 (0.00)	19624.81 (+5.98)

Another notable feature of QM/AMOEBA for these H-bonding systems is that it yields overly favorable polarization energies. To shed some light on that, we revisit the water dimer and the water-halide (F^-, Cl^-) complexes. Three different types of calculations are performed: (i) mutually polarizable; (ii) allowing the polarization of H-acceptor only; (iii) allowing the polarization of H-donor only, and their results are collected in Table 6.4. Here excluding the polarization of a certain fragment is realized by freezing the associated molecular orbitals or induced dipoles in an SCF calculation. (e.g., for an "H-donor only" ALMO-EDA (SCF-MI) calculation, the orbital rotations on the fragment corresponding to the H-acceptor are suppressed.) According to these results, when AMOEBA water serves as the H-donor, it consistently *under-polarizes* the H-acceptor in terms of the resulting polarization energy once its own polarization is forbidden, which should be related to its too weak electrostatic attraction with the electrons of the H-acceptor, as indicated in Table 6.3. It is the *over-polarization of AMOEBA water* as an H-donor (the so-called backward polarization) and the *exaggerated mutual polarization effect* that contribute to the overestimated polarization energies given by QM/AMOEBA for these systems. The most illustrative example is the water-F^- complex. Considering two unidirectional direct polarizations in QM/AMOEBA, the QM region (the H-acceptor F^-) is significantly *under-polarized* compared to the full QM reference based on

the energetic criterion, while the AMOEBA water is only *marginally over-polarized* by F^-. However, the mutual polarization effect (whose magnitude can be measured by the difference between the total polarization energy and the sum of two unidirectional polarization energies) in QM/AMOEBA is found to be substantially larger than that in full QM. We think that such an exaggerated mutual polarization effect is related to the lack of explicit treatment of Pauli repulsion that would otherwise curb excessive electric polarization in our QM/AMOEBA model. When the H-acceptor is also described by AMOEBA, the under-polarization of H-acceptor and over-polarization of H-donor both become more pronounced, and the resulting total polarization energy can be either less (water dimer, water-F^-) or more (water-Cl^-) favorable than that of QM/AMOEBA. Although for some systems such as the water dimer, the pure AMOEBA polarization energy is in good agreement with its QM counterpart, the underlying physical pictures are not in line with each other, as indicated by the relative strength of "H-donor only" and "H-acceptor only" polarization energies for the water dimer. Here we see that the AMOEBA water is excessively prone to electric polarization in general as a proton donor, and that the improper strength of mutual polarization effect due to the lack of explicit modeling of Pauli repulsion in the present QM/AMOEBA model further exacerbates the over-polarization problem.

Table 6.4: Polarization energies (in kcal/mol) for the water dimer and water-Cl^-, F^- complexes computed with ALMO-EDA and QM/AMOEBA, by allowing (i) mutual polarization, (ii) polarization of the H-acceptor only and (iii) polarization of the H-donor only. "D" (H-donor) and "A" (H-acceptor) are used to specify the regions described by QM or AMOEBA.

		mutual	H-acceptor only	H-donor only
water dimer	ALMO-EDA	-1.17	-0.73	-0.35
	AMOEBA	-1.23	-0.42	-0.64
	QM(A)/AMB(D)	-1.49	-0.56	-0.59
water-Cl^-	ALMO-EDA	-3.77	-0.99	-2.35
	AMOEBA	-6.83	-0.59	-5.52
	QM(A)/AMB(D)	-5.72	-0.85	-3.53
water-F^-	ALMO-EDA	-16.88	-5.33	-10.12
	AMOEBA	-19.08	-0.59	-16.67
	QM(A)/AMB(D)	-19.68	-2.35	-10.23

6.3.6 Convergence of the errors with the size of MM region

The last aspect that we investigate in this work is the convergence behavior of the errors demonstrated above with increasing sizes of the MM region, i.e., the number of AMOEBA water molecules surrounding the QM solute. Here we revisit three solutes that have been previously studied by us in Ref. 168: NH_3, NH_4^+ and CN^-, which form another set of

representatives of neutral, cationic and anionic species. For each solute, we choose one single snapshot from MD simulation (see Ref. 168 for the simulation details) and vary the number of solvent water molecules. By starting from the 10 water molecules that are closest to the solute (the position of the latter is marked by the center of the simulation sphere), we include 10 more water molecules that are the next closest at a time until the number of solvent molecules reaches 100, which is the maximum size of the MM region in this work. In order to make the full QM benchmarks for these systems computationally less demanding, we switch to the B97M-V functional [73] (which is a semi-local meta-GGA with the VV10 NLC) and the smaller def2-SVPD basis set. The choice of this basis has been validated by a benchmark study on systems containing 10–30 water molecules, where def2-SVPD and def2-TZVPPD yield similar full QM interaction energies once counterpoise corrections are applied (see Table E.1). Correspondingly, the same density functional and basis set are applied to the QM region in the QM/AMOEBA calculations for these systems.

With respect to the full QM references, the errors of QM/AMOEBA in total interaction energies and their components for three solutes surrounded by varying numbers of water molecules are shown in Figure 6.6 (the original data for interaction energies and EDA are provided in Table E.4). Note that the solute-solvent interaction energies (and the energy components) are not guaranteed to vary monotonically with the number of solvent molecules since those distant solvent molecules may not be aligned favorably relative to the solute. For each system, the error in total interaction energy converges to a certain value with increasing sizes of the MM region, and so do the errors in three energy components. This is reasonable because both polarization and vdW interaction (including exchange-repulsion, dispersion and other short-range effects such as charge transfer) decay fairly rapidly with respect to distance, and the multipole moments on those distant solvent molecules should be able to give a proper description for their long-range electrostatic interaction with the QM region.

Turning to the results for each individual solute, NH_3 and NH_4^+ show reasonably balanced QM/MM interfaces, for which the errors in permanent electrostatics are largely compensated by the sufficiently softened vdW potentials such that the deviations in total interaction energy are small in magnitude. Contrary to most of the systems investigated above, QM/AMOEBA *overbinds* NH_4^+ with the solvent water molecules, and the absolute error fortuitously decreases with the addition of more water shells until the number of H_2O molecules reaches 50, thanks to the faster-increasing error in permanent electrostatics that makes the interaction energy less overbound. For NH_3, QM/AMOEBA only slightly underestimates its interaction with the first 10 water molecules (by 0.15 kcal/mol), and the error increases to about 1.8 kcal/mol when the number of water molecules reaches 100. Also, with the addition of the first several water shells, we notice that the magnitude of errors in permanent electrostatics (and vdW interaction) varies more in the NH_4^+ case, indicating the difference between neutral and ionic solutes. CN^-, on the other hand, manifests the typical behavior of an anionic solute. The unbalanced errors in permanent electrostatics and vdW interactions result in the systematic underestimation of total interaction energy. It is also unusual that the errors in vdW interaction almost align with those in polarization energy. While a negative error in polarization energy whose magnitude is slightly larger than 5 kcal/mol

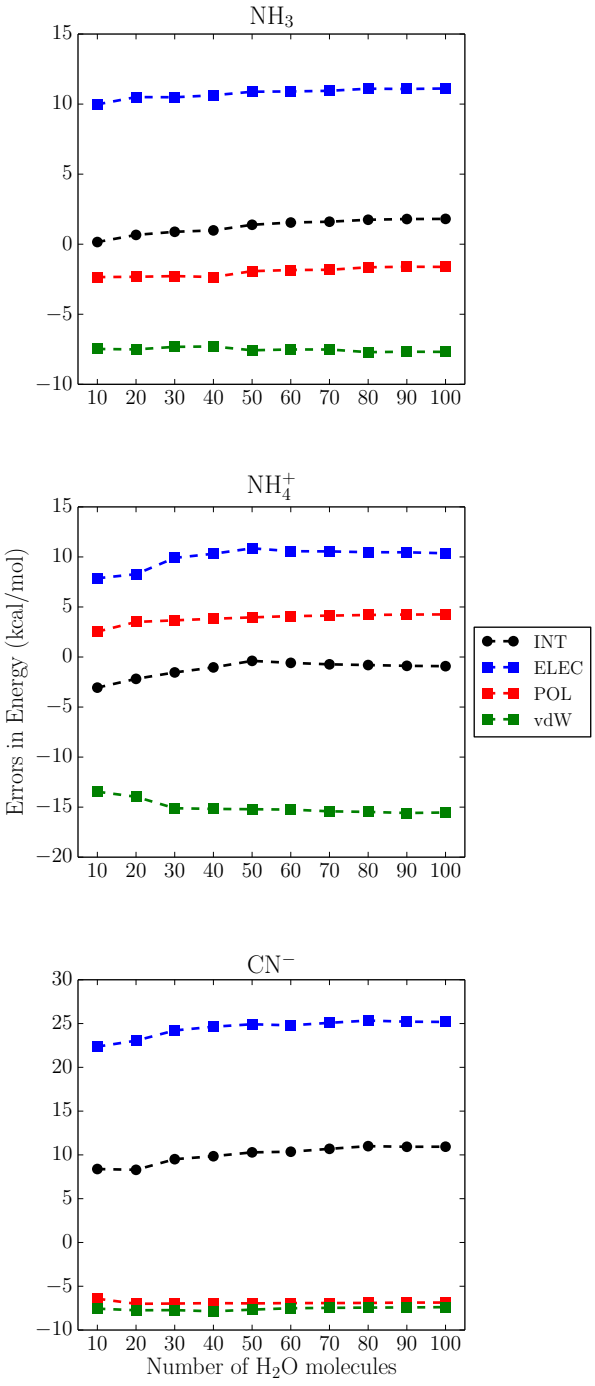

Figure 6.6: Errors of QM/AMOEBA (in kcal/mol) with respect to the full QM references for the solute-solvent interaction energies and their components when three solutes (NH_3, NH_4^+, and CN^-) are solvated with increasing numbers of water molecules. Here "QM" stands for the B97M-V/def2-SVPD level of theory.

is typical for anionic solutes (*vide supra*), the vdW potential here is undoubtedly too hard considering the significant charge penetration effect associated with CN^-. Nevertheless, in terms of the convergence behavior of the total interaction energy and its components, CN^- is *not* qualitatively different from the other two solutes.

6.4 Discussion and Outlook

The results shown in Figure 6.6 indicate that aside from the fortuitous error cancellation between different energy components, the errors in solute-solvent interaction energies given by the present QM/AMOEBA model primarily arise from the interaction with solvent molecules in the close vicinity, while the long-range portion of the QM/MM interaction seems to be properly described. For instance, considering the solute-solvent electrostatic interactions for the three species studied in Sec. 6.3.6, the interactions with the first 10 water molecules account for 75–90% of the errors at the bulk limit (which can be estimated by the nearly converged errors with increasing numbers of water shells). Therefore, in order to reduce the errors accumulated in the short range and thus improve the accuracy of the resulting solute-solvent interaction energies, it seems necessary to treat the solvent molecules in close proximity to the solute with more sophisticated models.

The conceptually simplest approach is to absorb these solvent molecules into the QM region, which, however, might significantly increase the computational demand. Meanwhile, it is often challenging to choose an appropriate size for the QM region *a priori*, as the convergence of QM/MM results with respect to the size of the QM region has been shown to be rather slow and not as monotonic as one might expect. [157–159, 514–517] In these studies, the size of the QM region required to reach the asymptotic limits for various properties often goes up to hundreds of atoms and sometimes exceeds the size of the largest model system in the present work. With the three systems discussed in Sec. 6.3.6, we evaluated the solute-solvent interaction energies with a new set of QM/AMOEBA calculations where the first 10 closest water molecules are also included into the QM region. The results indicate that for a given solute-solvent system, including some solvent molecules into the QM region based on proximity does not consistently improve the resulting solute-solvent interaction energy. We refer the readers to Appendix E.4 for a more detailed discussion.

Another plausible approach based on Figure 6.6 is to add a buffer layer in the middle of QM and MM regions wherein an improved AMOEBA model with modified functional forms is applied to describe the solvent molecules when we consider their interactions with the QM solute, while the interactions between solvent molecules (including those in the buffer layer) remain unchanged (still described with the original AMOEBA model). This buffer region is devised to reduce the errors in short-range solute-solvent interactions while providing a smooth transition from QM to AMOEBA, and methods such as adaptive QM/MM [518, 519] can be used to ensure the smoothness of the PES when solvent molecules move in and out of the buffer region. We note that a buffer layer similar to this has been used by Olsen et al. in the formulation of their polarizable density embedding (PDE) model, [495] wherein the

molecules have both frozen molecular orbitals (QM-like) and inducible dipoles (MM-like).

As a roadmap to a modified AMOEBA model whose interfacing with QM is improved, the first obvious goal is the proper treatment of charge penetration effect. There have been many efforts made aiming to incorporate this effect in various MM and QM/MM models, [153, 362–367, 390, 396, 479–484, 520–524] and under the framework of the AMOEBA force field, appropriate functional forms that account for charge penetration in the context of pure MM calculations have also been suggested in recent works.[142, 143] Many of these methods are similar in spirit: on each MM site, the point charge or point multipoles are first separated into nuclear and electron contributions (similar to what has been done for the data in Table 6.3), and then the latter is replaced by a continuous distribution (e.g. Slater or Gaussian function) with a certain spatial extent, or alternatively, damping functions are applied to the components of electrostatic interactions that involve electrons. The development of these models has provided many options that can be potentially adopted in an improved QM/AMOEBA model.

However, we find that although applying a scheme similar to that in Ref. 142 to our QM/AMOEBA model is able to improve the description of permanent electrostatics (we refer the readers to the left panel of Figure E.5 in Appendix E.4 and Figure 2 in Ref. 525 for preliminary results), it might exacerbate the so-called "electron-spill" effect, [526] i.e., the QM region is enormously over-polarized so that the electrons are pulled out of the QM region, and energetically it results in vastly overestimated polarization energies (see Figure E.6 and the right panel of Figure E.5). This is related to the aforementioned splitting of multipole moments which effectively places point charges of large magnitude (nuclei of MM atoms) near the QM/MM boundary, while the essential deficiency, nonetheless, is the purely empirical treatment of the repulsive vdW interaction (also termed exchange/Pauli/non-electrostatic repulsion), which fails to preclude the over-polarization of QM density by the electrostatic potential of MM. A model for Pauli repulsion that depends on electron density explicitly, despite being challenging because of the pure quantum nature of this interaction, is highly desirable not only for the sake of a more balanced and physically pertinent QM/MM interface, but also for the correct modeling of molecular properties that rely on the actual electronic structure, such as optical excitation energies. [526, 527] Popular methods for evaluating or incorporating Pauli repulsion in QM/MM (or other embedding) calculations include placing effective potentials on MM atoms [153, 154, 528–531] and other QM-derived approaches (e.g. the use of non-additive kinetic energy functionals, projection operators, etc.) that make direct use of pre-computed electron densities/MOs on embedding fragments [495, 532–539]. On the other hand, overlap-based models have also been proposed to mimic the effect of Pauli repulsion in pure MM [393, 394, 396, 540–546] and charge-dependent QM/MM [547, 548] models. Further investigation is required to seek for a suitable approach that explicitly accounts for the effect of Pauli repulsion between QM electrons and AMOEBA fragments with moderate computational costs.

6.5 Conclusion

In this work, we have presented a QM/polarizable MM model which employs modern density functionals and the AMOEBA force field. The total energy of the coupled QM/AMOEBA system is variationally minimized with respect to both the QM electron density and the AMOEBA induced dipoles following the procedure introduced in Sec. 6.2.1 so that the mutual polarization between QM and MM regions is treated in a fully self-consistent fashion. The implementation of this model is achieved through the Q-Chem/LibEFP code interface, where Q-Chem serves as the driver for the whole QM/AMOEBA calculation and LibEFP is modified to accommodate the additional functionalities for the support of AMOEBA. We note that LibEFP is a portable library so it can be readily interfaced with other quantum chemistry software packages.

The proposed QM/AMOEBA model is employed for the evaluation of interaction energies between several simple solutes (including neutral and ionic species) and various numbers of solvent water molecules, which are equivalent to the interactions between QM and MM regions. With the goal of investigating the source of errors in the resulting solute-solvent interaction energies, an EDA scheme is proposed to separate the total interaction energy crossing the QM/MM interface into contributions from permanent electrostatics, polarization and vdW interaction. This allows us to ascertain the agreement of each energy component with its counterpart in full QM calculations obtained by using the modified ALMO-EDA scheme. In general, the present QM/AMOEBA model yields reasonable total solute-solvent interaction energies for investigated neutral (H_2O, CH_4, and NH_3) and cationic (Na^+ and NH_4^+) species (often *not* for the correct reason, *vide infra*), but significantly underestimates the interactions for anionic solutes (Cl^-, F^-, and CN^-). Looking at the energy components more closely by means of EDA, the following points emerge:

- The permanent electrostatic interaction given by the current QM/AMOEBA model is always not sufficiently attractive, and it can suffer from even more significant CPE than in the pure AMOEBA scenario. This is most pronounced when the QM solute serves as the acceptor of protons from the MM region, as in systems like anion-water clusters.

- The vdW potential of AMOEBA is usually softened relative to its counterpart in full QM interaction energy. However, bearing in mind that the associated vdW parameters are fitted together with pure AMOEBA's permanent electrostatics and polarization, they might no longer be suitable for QM/AMOEBA whose these two energy components are both shifted (see Figures 6.2 and 6.3).

- The error in polarization energy is usually smaller compared to the discrepancies in other two energy components, and the sign of the error turns out to be system-dependent. According to our tests, QM/AMOEBA overestimates the polarization energy for H-bonding systems where AMOEBA water plays the role of a proton donor (which are roughly the same systems that have the largest CPEs), and it can substantially underestimate the polarization energy for other systems (such as $CH_4(H_2O)_{20}$) as well.

Taken together, as we have summarized at the end of Sec. 6.3.3, the quality of solute-solvent interaction energies given by the present QM/AMOEBA model strongly relies on the error cancellation amongst three components, especially the delicate balance between insufficiently favorable permanent electrostatics and vdW potential with extra softness. For several investigated solute-solvent systems, the "reasonable" total interaction energy masks the incorrectness of the underlying physics (one typical example is the $CH_4(H_2O)_{20}$ cluster), and such a "brittle" balance can break down once the permanent electrostatic and polarization components of QM/AMOEBA interactions are largely discrepant from their pure AMOEBA counterparts. It has also been shown that the errors in individual energy components are mostly rooted in the interaction with the solvent molecules in proximity to the QM region, and they converge rather rapidly with the addition of solvent shells, which implies that this model can be systematically improved by properly addressing the short-range discrepancy in each individual energy component. Future work will be devoted to this aspect based on the discussions in Sec. 6.4.

Appendix A

Additional Information for MAB-SCF

A.1 Additional Theoretical Details

A.1.1 More details about the gradient of the MAB objective function

Using the same notations as in Sec. 2.2.2, the most general form of the gradient of the MAB objective function is given by

$$\frac{\partial E}{\partial \Delta^{ZpZq}} = \sum_{X,Y} \frac{\partial R^{X\mu Y\nu}}{\partial \Delta^{ZpZq}} G_{Y\nu X\mu}, \tag{A.1}$$

where $\mathbf{G} = \mathbf{SPSRF} + \mathbf{FRSPS}$. With the parameterization of $\mathbf{B}$ by equations (2.14) and (2.15), we have

$$\begin{aligned}
\frac{\partial R^{X\mu Y\nu}}{\partial \Delta^{ZpZq}} &= \left[C^{X\mu}_{\ \ Xp}\delta^{X}_{\ Z}\sigma_{XqXi} - C^{X\mu}_{\ \ Xq}\delta^{X}_{\ Z}\sigma_{XpXi}\right](\sigma^{-1})^{XiYj}(B^{T})^{\ Y\nu}_{Yj} \\
&+ B^{X\mu}_{\ \ Xi}(\sigma^{-1})^{XiYj}\left[C^{Y\nu}_{\ \ Yp}\delta^{Y}_{\ Z}\sigma_{YqYj} - C^{Y\nu}_{\ \ Yq}\delta^{Y}_{\ Z}\sigma_{YpYj}\right] \\
&- \sum_{W}(RS)^{X\mu}_{\ \ W\sigma}\left[C^{W\sigma}_{\ \ Wp}\delta^{W}_{\ Z}\sigma_{WqWl} - C^{W\sigma}_{\ \ Wq}\delta^{W}_{\ Z}\sigma_{WpWl}\right](\sigma^{-1})^{WlYj}(B^{T})^{\ Y\nu}_{Yj} \\
&- \sum_{W}B^{X\mu}_{\ \ Xi}(\sigma^{-1})^{XiWl}\left[C^{W\sigma}_{\ \ Wp}\delta^{W}_{\ Z}\sigma_{WqWl} - C^{W\sigma}_{\ \ Wq}\delta^{W}_{\ Z}\sigma_{WpWl}\right](SR)^{\ Y\nu}_{W\sigma}.
\end{aligned} \tag{A.2}$$

and correspondingly the gradient becomes

$$\begin{aligned}
\frac{\partial E}{\partial \Delta^{ZpZq}} = &-2\sigma_{ZpZj}\left[(\sigma^{-1})B^{T}G(I-RS)C\right]^{Zj}_{\ \ Zq} \\
&+2\sigma_{ZqZj}\left[(\sigma^{-1})B^{T}G(I-RS)C\right]^{Zj}_{\ \ Zp}
\end{aligned} \tag{A.3}$$

When $\mathbf{\Delta}$ stands for the orbital rotations within the MAB space, i.e., $p = i$, $q = j$, matrix $\mathbf{C}$ reduces to $\mathbf{B}$, and the gradient vanishes because $(\mathbf{I} - \mathbf{RS})\mathbf{B} = \mathbf{0}$. On the other hand, orbital

rotations within the complementary space of the MAB ($p = a$, $q = b$) have no effect on the objective function value either, simply due to the enforced on-atom orthogonality ($\sigma_{Z_aZ_j} = 0$). Therefore, the only non-zero block of this gradient is resulted from the rotations between these two subspaces. If we set $p = i$, $q = a$, only the first term in Eq. (A.3) remains based on the arguments above, which immediately leads to the gradient represented by Eq. (2.19).

A.1.2 The preconditioned L-BFGS algorithm

The basic idea of L-BFGS is to construct the approximate Hessian (inverse Hessian in practice) for the current iteration with gradients and displacements computed in the most recent m steps, where m is the user-specified subspace size (number of "memorized" steps). If we define the gradient and displacement at kth iteration as $\mathbf{g_k}$ and $\mathbf{s_k}$, and $\mathbf{y_k} = \mathbf{g_{k+1}} - \mathbf{g_k}$, the kth approximate inverse Hessian can be evaluated as [256]:

$$\begin{aligned}\mathbf{H_k} = &\left(\mathbf{V_{k-1}^T}\cdots\mathbf{V_{k-m}^T}\right)\mathbf{H_k^0}\left(\mathbf{V_{k-m}}\cdots\mathbf{V_{k-1}}\right)\\ &+\boldsymbol{\rho}_{\mathbf{k-m}}\left(\mathbf{V_{k-1}^T}\cdots\mathbf{V_{k-m+1}^T}\right)\mathbf{s_{k-m}s_{k-m}^T}\left(\mathbf{V_{k-m+1}}\cdots\mathbf{V_{k-1}}\right)\\ &+\boldsymbol{\rho}_{\mathbf{k-m+1}}\left(\mathbf{V_{k-1}^T}\cdots\mathbf{V_{k-m+2}^T}\right)\mathbf{s_{k-m+1}s_{k-m+1}^T}\left(\mathbf{V_{k-m+2}}\cdots\mathbf{V_{k-1}}\right)\\ &+\cdots+\boldsymbol{\rho}_{\mathbf{k-1}}\mathbf{s_{k-1}s_{k-1}^T}.\end{aligned} \tag{A.4}$$

where

$$\boldsymbol{\rho}_{\mathbf{k}} = \frac{1}{\mathbf{y_k^T s_k}}, \quad \mathbf{V_k} = \mathbf{I} - \rho_{\mathbf{k}}\mathbf{y_k s_k^T}. \tag{A.5}$$

In practice, a "two-loop" algorithm which only requires evaluating vector-vector products is implemented to compute $\mathbf{H_k}$ acting on $\mathbf{g_k}$. $\mathbf{H_k^0}$ is the preconditioner for the approximate inverse Hessian. By default, a constant scaling factor is used for $\mathbf{H_k^0}$, which is considered as the unpreconditioned case here. Once $\mathbf{H_k^0}$ contains more information about the true inverse Hessian, the step generated by L-BFGS becomes closer to a Newton step, which can presumably accelerate the convergence.

If we still denote $\mathbf{G} = \mathbf{SPSRF} + \mathbf{FRSPS}$, the Hessian of the MAB objective function (Eq. (2.10)) can be formally represented as follows (only on-block mixings are allowed):

$$\begin{aligned}H_{X_iX_a,Y_jY_b} &= \left.\frac{\partial^2 E}{\partial\Delta^{X_iX_a}\partial\Delta^{Y_jY_b}}\right|_{\Delta=0}\\ &=\mathbf{R^{\Delta}SPSR^{\Delta}F} + \mathbf{R^{\Delta\Delta}G}\\ &=2(P_{ov})_{X_iY_b}(F_{ov})_{Y_jX_a} + 2(F_{ov})_{X_iY_b}(P_{ov})_{Y_jX_a}\\ &+2(P_{oo})_{X_iY_j}(F_{vv})_{Y_bX_a} + 2(F_{oo})_{X_iY_j}(P_{vv})_{Y_bX_a}\\ &-2(S_{ov})_{X_iY_b}(G_{ov})_{Y_jX_a} - 2(G_{ov})_{X_iY_b}(S_{ov})_{Y_jX_a}\\ &+2(S_{oo})_{X_iY_j}(G_{vv})_{Y_bX_a} - 2(G_{oo})_{X_iY_j}(S_{vv})_{Y_bX_a},\end{aligned} \tag{A.6}$$

where $\mathbf{R^{\Delta}}$ and $\mathbf{R^{\Delta\Delta}}$ stands for first- and second-order derivatives of $\mathbf{R}$ with respect to $\boldsymbol{\Delta}$. The first four terms on the RHS of Eq. (A.6) come from the "$\mathbf{R^{\Delta}R^{\Delta}}$" term, while the rest

from the "$\mathbf{R^{\Delta\Delta}}$" term. The explicit forms of the involved matrix elements are:

$$\begin{aligned}
(P_{oo})_{XiYj} &= \sigma_{XiXk}\left[\sigma^{-1}B^{T}SPSB\sigma^{-1}\right]^{XkYl}\sigma_{YlYj} \\
(P_{ov})_{XiYa} &= \sigma_{XiXj}\left[\sigma^{-1}B^{T}SPS(I-RS)\right]^{Xj}_{\ \ Y\nu}V^{Y\nu}_{\ \ Ya} \\
(P_{vv})_{XaYb} &= (V^{T})^{X\mu}_{\ \ Xa}\left[(I-SR)SPS(I-RS)\right]_{X\mu Y\nu}V^{Y\nu}_{\ \ Yb},
\end{aligned} \tag{A.7}$$

$$\begin{aligned}
(F_{oo})_{XiYj} &= \sigma_{XiXk}\left[\sigma^{-1}B^{T}FB\sigma^{-1}\right]^{XkYl}\sigma_{YlYj} \\
(F_{ov})_{XiYa} &= \sigma_{XiXj}\left[\sigma^{-1}B^{T}F(I-RS)\right]^{Xj}_{\ \ Y\nu}V^{Y\nu}_{\ \ Ya} \\
(F_{vv})_{XaYb} &= (V^{T})^{X\mu}_{\ \ Xa}\left[(I-SR)F(I-RS)\right]_{X\mu Y\nu}V^{Y\nu}_{\ \ Yb},
\end{aligned} \tag{A.8}$$

$$\begin{aligned}
(G_{oo})_{XiYj} &= \sigma_{XiXk}\left[\sigma^{-1}B^{T}GB\sigma^{-1}\right]^{XkYl}\sigma_{YlYj} \\
(G_{ov})_{XiYa} &= \sigma_{XiXj}\left[\sigma^{-1}B^{T}G(I-RS)\right]^{Xj}_{\ \ Y\nu}V^{Y\nu}_{\ \ Ya} \\
(G_{vv})_{XaYb} &= (V^{T})^{X\mu}_{\ \ Xa}\left[(I-SR)G(I-RS)\right]_{X\mu Y\nu}V^{Y\nu}_{\ \ Yb},
\end{aligned} \tag{A.9}$$

and

$$\begin{aligned}
(S_{oo})_{XiYj} &= \sigma_{XiXk}(\sigma^{-1})^{XkYl}\sigma_{YlYj} \\
(S_{vv})_{XaYb} &= (V^{T})^{X\mu}_{\ \ Xa}\left[S(I-RS)\right]_{X\mu Y\nu}V^{Y\nu}_{\ \ Yb} \\
(S_{ov})_{XiYa} &= \sigma_{XiXj}\left[(\sigma^{-1})B^{T}S\right]^{Xj}_{\ \ Y\nu}V^{Y\nu}_{\ \ Ya}.
\end{aligned} \tag{A.10}$$

In practice, we also have $\sigma_{XiXj} = \delta_{ij}$ since on-block orthogonality is enforced. More details about the Hessian derivation can be found in Ref. 117, which carefully derived the orbital Hessian for SCF-MI (a rather similar optimization problem).

The preconditioner we apply to the L-BFGS algorithm is the inverted on-diagonal blocks of the Hessian, i.e., the inverse of $\mathbf{H_{xx}}$ for all the different atom blocks (X). Within the "two-loop" implementation, $\mathbf{H_k^0}$ acts on vector $\mathbf{v} = \mathbf{V_{k-m}V_{k-m+1}}\cdots\mathbf{V_{k-1}g_k}$, which can be divided into contributions from each atom block. Therefore, the application of the preconditioner is equivalent to solving the following linear equation on each atom block:

$$H_{XiXa,XjXb}u^{XjXb} = v_{XiXa}, \tag{A.11}$$

where $\mathbf{u}$ is the preconditioned vector: $\mathbf{u} = \mathbf{H}_k^0\mathbf{v}$. Based on the property of the Hessian matrix (symmetric positive-definite), a preconditioned conjugate-gradient (CG) algorithm is implemented to solve Eq. (A.11) iteratively on each atom block. The implemented preconditioner for CG is actually the inverse of the on-diagonal part ($X = Y$) of the last two terms in Eq. (A.6).

A.1.3 Construction of pseudo-canonicalized MOs upon MAB-SCF solution

Once MAB-SCF converges, a Fock matrix in the secondary basis can be built upon the MAB density matrix projected into the secondary basis:

$$\mathbf{F} = \mathbf{F}(\tilde{\mathbf{P}}), \quad \tilde{\mathbf{P}} = \mathbf{B}\mathcal{P}\mathbf{B}^T. \tag{A.12}$$

In the current implementation, the PT2 correction is evaluated based on pseudo-canonicalized occupied and virtual MOs, which can be obtained by diagonalizing $\mathbf{F}_{OO}$ and $\mathbf{F}_{VV}$ separately. In fact, the occupied ones are already available in this case, since we can simply project the occupied MOs optimized by MAB-SCF into the secondary basis: $(C_o)^{\mu}_{\ i} = B^{\mu}_{\ \alpha}(\mathcal{C}_o)^{\alpha}_{\ i}$. Obviously, $\mathbf{C}_o$ diagonalizes $\mathbf{F}$:

$$(C_o^T)_i^{\ \mu} F_{\mu\nu} (C_o)^{\nu}_{\ j} = (\mathcal{C}_o^T)_i^{\ \alpha} F_{\alpha\beta} (\mathcal{C}_o)^{\beta}_{\ j} = \epsilon_i \delta_{ij}. \tag{A.13}$$

To obtain the eigenvalues and eigenvectors of $\mathbf{F}_{VV}$, we first form an orthonormal basis that spans the virtual space. If the full but non-redundant span of the secondary basis is represented by $\mathbf{X}$ ($\mathbf{X}^T\mathbf{S}\mathbf{X} = \mathbf{I}$), the demanded vectors can be generated by projecting out the space spanned by occupied MOs:

$$\mathbf{V} = (\mathbf{I} - \tilde{\mathbf{P}}\mathbf{S})\mathbf{X}. \tag{A.14}$$

The vectors in $\mathbf{V}$ can be orthonormalized again by performing a canonical orthogonalization (diagonalizing $\mathbf{V}^T\mathbf{S}\mathbf{V}$ will be required). Also, after doing this, the linear dependency of vectors in $\mathbf{V}$ will be eliminated and its column dimension reduces to N_v. We denote the resulting orthonormal basis as $\mathbf{V}'$. Solving the following standard eigenvalue problem

$$(\mathbf{V}'^T\mathbf{F}\mathbf{V}')\mathbf{C}'_v = \boldsymbol{\epsilon}_v \mathbf{C}'_v, \tag{A.15}$$

the energies of the pseudo-canonicalized virtual orbitals (ϵ_a's in Eq. 2.29) are given by $\boldsymbol{\epsilon}_v$, and their coefficients $\mathbf{C}_v = \mathbf{V}'\mathbf{C}'_v$. The Fock matrix elements coupling between occupied and virtual pseudo-canonicalized MOs ($\mathbf{F}_{OV}$) can be evaluated as

$$F^{(1)}_{ia} = (\mathcal{C}_o^T)_i^{\ \alpha} (B^T)_{\alpha}^{\ \mu} F_{\mu\nu} (C_v)^{\nu}_{\ a}. \tag{A.16}$$

A.2 Additional Benchmark Results

A.2.1 Effect of the "mixing" parameter on the performance of MAB-SCF (PC) for small anions

Since the "mixing" parameter determines the shift applied to all the eigenvalues of the Fock matrix (orbital energies), it modifies the objective function (Eq. 24 in the main text) directly and thus changes the optimized MAB. To see how the results of MAB-SCF (PC) for anions are affected by this parameter, we tested $\alpha = 0.25, 0.50, 0.75, 1.00$ with the G21EA dataset (electron affinities for 21 monoatomic and diatomic species).

Table A.1: RMSDs (in kcal/mol) of MAB-SCF (PC) for the G21EA dataset with different values for the "mixing" parameter: $\alpha = 0.25, 0.50, 0.75, 1.00$, compared to the results of conventional SCF. The computational details are the same as in Table 2.7 in Chap. 2 (the "adding vector" strategy is applied for NO^-, PO^-, O_2^-, and S_2^-).

α value	MAB-SCF			MAB-SCF (PC)		
	B97-D	B97M-V	B3LYP	B97-D	B97M-V	B3LYP
0.25	2.506	2.599	2.705	0.269	0.191	0.034
0.50	2.508	2.597	2.705	0.151	0.065	0.031
0.75	1.940	2.082	2.221	0.101	0.027	0.078
1.00	2.522	2.602	2.721	0.078	0.037	0.137

Based on the results in Table A.1, MAB-SCF is most accurate when $\alpha = 0.75$ for all three tested functionals, which gives the main reason why we choose that number in Chap. 2. The optimal α that gives the smallest RMSD after PC is applied, on the other hand, varies for different functionals: while larger α helps reduce the error for the tested GGA functional (B97-D), smaller value is preferred by the hybrid (B3LYP). This difference is physically reasonable, because for a given system, the calculated HOMO-LUMO gap in principle should increase in the B97-D<B97M-V<B3LYP order. In the present work, we simply choose $\alpha = 0.75$ as a compromise for all functionals, when positive ϵ(HOMO) calculated by PRB-SCF occurs. Although it is possible to fine-tune the α value for different categories of density functionals, we use this uniform value for the entire paper since these monoatomic or diatomic anions are *not* deemed as the main target of MAB-SCF (PC) after all, because of their tiny sizes.

A.2.2 Effect of choosing different reference bases (RB)

The reference density matrix upon which MAB is fitted is provided by a converged solution of PRB-SCF, and the rank of the PRB is determined by the (unprojected) reference

basis (RB) — usually a standard basis set of double-ζ quality. Using different RBs will affect the quality of fitted MAB: while using a larger reference basis, in principle, should be able to give an MAB whose quality is closer to that of PAO, the associated computational cost for the PRB-SCF steps also increases. Therefore, a compromise between accuracy and efficiency is needed.

Table A.2: Errors (in kcal/mol) of MAB-SCF (PC) for the "pruned" (hypervalent molecules excluded) and full G2 set with using four different RBs: 6-31G, 6-31+G, 6-31G(d), 6-31+G(d), compared to the results of conventional SCF in aQZ (the secondary basis set employed in the MAB calculations). The B97M-V functional is used for all the calculations. The "adding vector" strategy is used for the same group of hypervalent molecules as specified in Sec. 2.4.1.

	Pruned G2 set			Full G2 set		
Ref basis	MAX	RMSD	MSE	MAX	RMSD	MSE
6-31G	-0.78	0.23	-0.16	-1.03	0.28	-0.18
6-31+G	-0.84	0.20	-0.13	-0.94	0.24	-0.14
6-31G(d)	-0.21	0.06	-0.04	-0.21	0.06	-0.04
6-31+G(d)	-0.14	0.03	-0.02	-0.14	0.03	-0.02
PAO	-0.08	0.02	-0.01	–	–	–

Table A.2 shows how the performance of MAB-SCF (PC) on the G2 set depends on the choice for the reference basis. Without considering the hypervalent molecules, we see that including one set of polarization functions in the double-ζ Pople basis significantly improves the quality of the fitted MAB, and that one set of extra diffuse functions ("+") also helps reduce the error by roughly a constant amount. With hypervalent molecules included, the improvement brought by adding the polarization functions becomes even larger, which indicates the necessity of including one set of d functions in the basis set to reasonably describe the electronic structure of these species. It should be noted that it is not guaranteed that the error will be reduced on each individual data point when switching to a larger RB (e.g. the size of the maximum error on the pruned G2 set increases when changing the RB from 6-31G to 6-31+G), mostly due to the characteristics of PC. Nevertheless, the general trend should be consistent with that revealed by Table A.2, since a larger RB provides a better initial description of the chemical environment. In this work, we choose the relatively larger 6-31+G(d) basis, and the resulting MAB has fairly comparable accuracy to that of the PAO.

A.2.3 Using fully environment-adapted dimensions for the MAB

It has been argued that one limitation of this work is the "semi-automated" modification of the MAB dimensions, i.e. the algorithm that increases the sizes of the MAB based on

the reference density matrix from PRB-SCF is only applied to user-specified hypervalent molecules. It would be more ideal if such a procedure (and the "*thresh*2" parameter we choose) could be applied to all the systems without any further change, which means, we have fully environment-adapted dimensions for the MAB.

Table A.3: Errors (in kcal/mol) of MAB-SCF (PC) for the G2 set with the "adding vector" algorithm applied to (i) hypervalent molecules only and (ii) all species, compared to the results of conventional SCF in aQZ (secondary basis for the MAB-SCF calculations). PT2 and DFPC are used for pure (B97-D, B97M-V) and hybrid (B3LYP, M06-2X, ωB97X-V) functionals, respectively.

	Hypervalent only			All		
	MAX	RMSD	MSE	MAX	RMSD	MSE
B97-D	-0.26	0.05	-0.02	-0.26	0.03	0.01
B97M-V	-0.14	0.03	-0.02	-0.14	0.03	0.01
B3LYP	0.33	0.08	0.05	0.33	0.03	0.01
M06-2X	0.26	0.08	0.06	0.20	0.04	0.03
ωB97X-V	0.55	0.13	0.09	0.34	0.06	0.04

Such a test is performed on the G2 set and the results are collected in Table A.3. In contrast to the scheme adopted in Chap. 2 ("Hypervalent only"), applying the "adding vector" algorithm to all the species reduces the RMSDs for the tested functionals exclusively, which indicates that the MAB dimensions for part of the molecules that are previously designated as "non-hypervalent" are also augmented by the same algorithm. This is preferable in terms of accuracy, especially for hybrid functionals where the RMSDs on the full G2 set are almost halved. However, as we have mentioned, the default value for *thresh*2 used in this work is tuned to obtain satisfactory accuracy for the *hypervalent molecules only*, thus it is likely to add an excessive number of extra functions into the MAB for other molecules, which makes the MAB optimization problem more challenging to converge, possibly caused by redundant degrees of freedom.

Table A.4 demonstrates how the MAB dimensions and number of iterations consumed to converge the optimization problem change for several "non-hypervalent" molecules in the G2 set. While no change occurs on those simple compounds that are almost "absolutely non-hypervalent" (the top half), the sizes of the MAB could increase by a large amount for molecules containing π-conjugated systems or those with largely polar (like carbon-halogen) covalent bonds. On one hand, expanding the MAB sizes for these molecules could help improve the accuracy of MAB-SCF (PC) on these systems, since they turn out to be the most difficult cases for a minimal-rank basis; on the other hand, we believe that the current "adding vector" approach is *arguably not ideal* yet to be applied onto all generic molecules. For instance, we would expect that C_2F_4 and C_2Cl_4 have roughly the same number of new functions added because of their similar electronic structure. In reality, while 3 functions are

added onto each carbon atom in C_2F_4 (6 new functions in total), in C_2Cl_4 4 are added onto each carbon and 2 onto each chlorine (16 new functions in total). The treatment of these two molecules are rather uneven and the excessive size of the expanded MAB for C_2Cl_4 makes its optimization about 8 times more difficult to converge. Moreover, the cost for MAB-SCF also gets increased if the size of the MAB is largely expanded: in the case of C_6H_6 and C_2Cl_4, the number of MAB functions is increased by about 1/3. Based on these reasons, we are not inclined to apply the current "adding vector" algorithm that is mainly developed for hypervalent molecules to other systems without further development.

Table A.4: Illustration of changes in the MAB dimensions and number of iterations required to converged the MAB optimization problem for several "non-hypervalent" G2 molecules before and after applying the "adding vector" algorithm. Data are collected with using the B97M-V functional and aQZ as the secondary basis.

	without add_vec		with add_vec	
	rank (MAB)	Iter count	rank (MAB)	Iter count
CH_4	9	96	9	96
H_2O	7	89	7	89
CO_2	15	75	15	75
C_2H_4	14	121	14	121
C_6H_6	36	119	48	494
CF_4	25	36	32	231
CCl_4	41	34	49	153
C_2F_4	30	70	36	198
C_2Cl_4	46	80	62	653

A.2.4 Comparison of MAB-SCF (PC) against dual-basis SCF

The dual-basis (DB)-SCF method uses an exact subset of the target secondary basis (prepared by truncation or projection) as the primary basis in which an SCF solution is computed. After that, a single Roothaan step is taken in the secondary basis in order to approach the SCF energy evaluated in it. It turns out to be a very useful method to reduce the basis set incompleteness error with lower computational cost. In MAB-SCF (PC), the SCF solution is computed using the MAB (an exact subset of the secondary basis) and then further corrected by PT2 or DFPC, which is rather similar to the procedure of DB-SCF. Here we briefly compare the accuracy of MAB-SCF (PC) and DB-SCF when measured against the exact SCF solutions.

Table A.5 shows the RMSDs of DB-SCF and MAB-SCF (PC) for the G2 set, where aug-cc-pVQZ is used as the common secondary basis. DB-SCF demonstrates very good accuracy

Table A.5: RMSDs of DB-SCF (with different basis parings) and MAB-SCF (PC) for the G2 set evaluated with three density functionals, measured against the results computed by conventional SCF. aQZ is employed as the secondary basis for all the calculations. In the first two DB parings, the primary basis is prepared by projecting the small basis (6-31+G(d), raTZ) into the secondary basis to make the former an exact basis subset of the latter.

DB pairing	B97-D	B97M-V	B3LYP
aQZ/6-31+G(d)	0.276	0.613	0.169
aQZ/raTZ	0.096	0.065	0.059
aQZ/raQZ	0.066	0.053	0.051
MAB-SCF (PC)	0.053	0.033	0.083

with reduced-aug-cc-pVQZ (raQZ), the fine-pruned basis subset of aQZ (see Ref. 215). It should be noted that the size of raQZ is only marginally smaller than aug-cc-pVTZ, therefore, to calculate an SCF solution in it might still be rather expensive for large systems. To see how the performance of DB-SCF relies on the paired primary basis, we prepared two other primary basis sets by projecting 6-31+G(d) and reduced-aug-cc-pVTZ (raTZ, which is slightly larger than aDZ) into the secondary basis. The procedure is exactly the same as how the PRB is constructed in MAB-SCF (PC). In the projected raTZ case, the RMSDs only slightly increase, while the errors associated with the "aQZ/6-31+G(d) (projected)" pairing grow by 3–10 times compared to the aQZ/raQZ case for different functionals. Turning to the results of MAB-SCF (PC), in general it provides comparable accuracy to DB-SCF with aQZ/raQZ basis paring: it slightly outperforms for the two tested pure functionals (B97-D, B97M-V) while slightly losses on the hybrid functional (B3LYP), which is consistent with the strength and weakness of MAB-SCF (PC) revealed by the previous assessments.

Since these available Dunning basis subsets were originally trained on the S22 dataset (Ref. 67), it would also be interesting to compare the performance of DB-SCF and MAB-SCF (PC) on datasets for non-covalent interactions (NC) as well. Using aug-cc-pVTZ (aTZ) as the secondary basis (mainly because there is no fine-tuned basis subset available for def2-QZVPPD), Table A.6 demonstrates the RMSDs of DB-SCF and MAB-SCF (PC) for the six NC datasets that we benchmarked on in Chap. 2. With the properly trained raTZ basis (which is slightly larger than aDZ) paired with aTZ, DB-SCF does show superior accuracy compared to that of MAB-SCF (PC) on most of these NC datasets except for the binding energies of $F^-(H_2O)_{10}$ isomers. This is not surprising because in fact raTZ and aTZ only differ by one diffuse d function and two f (one of which is diffuse) functions, which might not be the most important components to capture weak intermolecular interactions. It is noteworthy that with roughly 1.5× size of the primary basis, raTZ reduces the DB-SCF error by almost an order of magnitude compared to the projected 6-31+G(d) basis. The accuracy of MAB-SCF (PC), on the other hand, is still quite close to that given by DB-

Table A.6: RMSDs of DB-SCF (with different basis pairings) and MAB-SCF (PC) for 6 NC datasets (the same ones as we benchmarked in Chap. 2), compared against the results of conventional SCF evaluated in aTZ (the secondary basis in DB and MAB calculations). All the binding energies are calculated using the B97M-V functional. The primary basis in "DB-SCF (aTZ/6-31+G(d))" is prepared by projecting 6-31+G(d) into the aTZ basis.

	DB-SCF (aTZ/6-31+G(d))	DB-SCF(aTZ/raTZ)	MAB-SCF (PC)
A24	0.057	0.004	0.007
S22	0.064	0.008	0.017
HB15	0.134	0.010	0.024
H2O6Bind8	0.871	0.066	0.091
FmH2O10	1.223	0.129	0.114
L7	0.095	0.036	0.044

SCF (aTZ/raTZ), and the largest difference (0.025 kcal/mol) appears on the dataset for the binding energies of water hexamers.

Appendix B

Additional Information for the "Adiabatic" EDA

B.1 Extra Details about the Derivation of $E^x_{\mathbf{frz}}$

Notations used in the equations presented in Secs. B.1.1–B.1.3: Greek letters μ, ν, λ, ...: AO basis indices; lowercase Romans i, j, k, ...: occupied MO indices; a, b, c, ...: virtual MO indices; p, q, r, ...: generic MO indices; uppercase Romans X, Y, Z, ...: fragment indices for ALMOs. The equations are derived with spin-orbitals (α or β) so that the resulting equations can be applied to both restricted and unrestricted calculations. Tensorial notations are used throughout, i.e., superscripts refer to contravariant indices while subscripts refer to the covariant ones. Einstein summation convention is applied to contractions between contravariant and covariant indices, except for summations over indices on different fragments, which will be shown explicitly.

B.1.1 Derivation of $E^{\mathbf{S}}_{\mathbf{frz}} \cdot \mathbf{S}^x$

Based on the expression of the initial density: $P^{A\lambda B\sigma}_{\text{frz}} = (C_o)^{A\lambda}_{\ Ai}(\sigma^{-1})^{AiBj}(C_o^T)^{\ B\sigma}_{Bj}$, we have

$$\sum_{YZ}\frac{\partial E_{\text{frz}}}{\partial S_{Y\mu Z\nu}}\cdot S^x_{Y\mu Z\nu} = \sum_{ABYZ}\frac{\partial E_{\text{frz}}}{\partial P^{A\lambda B\sigma}_{\text{frz}}}\cdot\frac{\partial P^{A\lambda B\sigma}_{\text{frz}}}{\partial S_{Y\mu Z\nu}}\cdot S^x_{Y\mu Z\nu} \tag{B.1}$$

$$\begin{aligned}
&= \sum_{ABYZ} F_{A\lambda B\sigma}\left[\frac{\partial (C_o)^{A\lambda}_{\ Ai}}{\partial S_{Y\mu Z\nu}}(\sigma^{-1})^{AiBj}(C_o)^{\ B\sigma}_{Bj} + (C_o)^{A\lambda}_{\ Ai}(\sigma^{-1})^{AiBj}\frac{\partial (C_o)^{\ B\sigma}_{Bj}}{\partial S_{Y\mu Z\nu}}\right.\\
&\quad\left. + (C_o)^{A\lambda}_{\ Ai}\frac{\partial (\sigma^{-1})^{AiBj}}{\partial S_{Y\mu Z\nu}}(C_o)^{\ B\sigma}_{Bj}\right]\cdot S^x_{Y\mu Z\nu}\\
&= \sum_{ABYZ} F_{A\lambda B\sigma}\left[\frac{\partial (C_o)^{A\lambda}_{\ Ai}}{\partial S_{Y\mu Z\nu}}(\sigma^{-1})^{AiBj}(C_o)^{\ B\sigma}_{Bj} + (C_o)^{A\lambda}_{\ Ai}(\sigma^{-1})^{AiBj}\frac{\partial (C_o)^{\ B\sigma}_{Bj}}{\partial S_{Y\mu Z\nu}}\right.\\
&\quad\left. - (C_o)^{A\lambda}_{\ Ai}\sum_{CD}(\sigma^{-1})^{AiCk}\frac{\partial \sigma_{CkDl}}{\partial S_{Y\mu Z\nu}}(\sigma^{-1})^{DlBj}(C_o)^{\ B\sigma}_{Bj}\right]\cdot S^x_{Y\mu Z\nu}
\end{aligned} \tag{B.2}$$

$\boldsymbol{\sigma}$ is the metric of all the occupied orbitals: $\sigma_{CkDl} = (C_o)^{C\alpha}_{\ Ck}S_{C\alpha D\beta}(C_o)^{D\beta}_{\ Dl}$. Therefore,

$$\frac{\partial \sigma_{CkDl}}{\partial S_{Y\mu Z\nu}} = \frac{\partial (C_o)^{C\alpha}_{\ Ck}}{\partial S_{Y\mu Z\nu}}S_{C\alpha D\beta}(C_o)^{D\beta}_{\ Dl} + (C_o)^{C\alpha}_{\ Ck}S_{C\alpha D\beta}\frac{\partial (C_o)^{D\beta}_{\ Dl}}{\partial S_{Y\mu Z\nu}} + (C_o)^{C\alpha}_{\ Ck}\delta^{Y\mu}_{C\alpha}\delta^{Z\nu}_{D\beta}(C_o)^{D\beta}_{\ Dl} \tag{B.3}$$

The occupied orbitals on fragment A can be parameterized as

$$(C_o)^{A\lambda}_{\ Ai} = C^{A\lambda}_{\ Ar}(\sigma_{AA}^{-\frac{1}{2}})^{Ar}_{\ Ap}[\delta^{Ap}_{\ Ai} + \Delta^{ApAq}\sigma_{ApAi} - \Delta^{AqAp}\sigma_{AqAi}] \tag{B.4}$$

where the $\boldsymbol{\sigma}_{AA}^{-\frac{1}{2}}$ term enforces on-fragment orthogonality. Thus, we have

$$\begin{aligned}
\left.\frac{\partial (C_o)^{A\lambda}_{\ Ai}}{\partial S_{Y\mu Z\nu}}\right|_{\boldsymbol{\sigma}_{AA}=1,\boldsymbol{\Delta}^{AA}=0} &= -\frac{1}{2}C^{A\lambda}_{\ Ar}\frac{\partial (\sigma_{AA})_{ArAp}}{\partial S_{Y\mu Z\nu}}\delta^{Ap}_{\ Ai}\\
&= -\frac{1}{2}C^{A\lambda}_{\ Ar}C^{\ A\pi}_{Ar}\delta^{Y\mu}_{A\pi}\delta^{Z\nu}_{A\rho}C^{A\rho}_{\ Ap}\delta^{Ap}_{\ Ai}\\
&= -\frac{1}{2}(S^{-1}_{AA})^{A\lambda A\pi}\delta^{Y\mu}_{A\pi}\delta^{Z\nu}_{A\rho}(C_o)^{A\rho}_{\ Ai}
\end{aligned} \tag{B.5}$$

Substituting Eq.(B.3) into Eq.(B.2) then using Eq. (B.5), we get

$$\begin{aligned}
&\sum_{YZ} \frac{\partial E_{\mathrm{frz}}}{\partial S_{Y\mu Z\nu}} \cdot S^x_{Y\mu Z\nu} \\
=& -\frac{1}{2}\sum_{A}(PF)^{A\rho}_{\ \ A\lambda}(S^{-1}_{AA})^{A\lambda A\pi} \cdot S^x_{A\pi A\rho} \\
&-\frac{1}{2}\sum_{B}(S^{-1}_{BB})^{B\pi B\sigma}(FP)^{\ \ B\rho}_{B\sigma} S^x_{B\rho B\pi} \\
&+\frac{1}{2}\sum_{C}(S^{-1}_{CC})^{C\pi C\alpha}(SPFP)^{\ \ C\rho}_{C\alpha} S^x_{C\rho C\pi} \\
&+\frac{1}{2}\sum_{D}(PFPS)^{D\rho}_{\ \ D\beta}(S^{-1}_{DD})^{D\beta D\pi} \cdot S^x_{D\pi D\rho} - \sum_{CD}(PFP)^{D\beta C\alpha} S^x_{C\alpha D\beta} \\
=& -\frac{1}{2}\sum_{A}\left\{(\mathbf{S}^{-1}_A)[(\mathbf{I}-\mathbf{SP})\mathbf{FP}]_A + [\mathbf{PF}(\mathbf{I}-\mathbf{PS})]_A(\mathbf{S}^{-1}_A)\right\} \cdot \mathbf{S}^x_A - \mathbf{W}\cdot\mathbf{S}^x
\end{aligned} \tag{B.6}$$

which gives Eq. (3.15).

B.1.2 Evaluation of the fragment response gradient

The z-vector equation

We start from the stationary condition of fragment A's SCF solution: $E_A^{\Delta_A} = 0$. Therefore, we have

$$\frac{\mathrm{d}}{\mathrm{d}x} E_A^{\Delta_A} = 0 \tag{B.7}$$

which leads to

$$\begin{aligned}
&E_A^{\Delta_A\mathbf{S}_A}\cdot\mathbf{S}^x_A + E_A^{\Delta_A\mathbf{h}_A}\cdot\mathbf{h}^x_A + E_A^{\Delta_A\mathrm{II}_A}\cdot\mathrm{II}^x_A + E^{\Delta_A x}_{\mathrm{xc},A} + E_A^{\Delta_A\Delta_A}\cdot\Delta^x_A = 0 \\
\Longrightarrow\ & E_A^{\Delta_A\Delta_A}\cdot\Delta^x_A = -(E_A^{\Delta_A\mathbf{S}_A}\cdot\mathbf{S}^x_A + E_A^{\Delta_A\mathbf{h}_A}\cdot\mathbf{h}^x_A + E_A^{\Delta_A\mathrm{II}_A}\cdot\mathrm{II}^x_A + E^{\Delta_A x}_{\mathrm{xc},A}) \qquad \text{(B.8)} \\
\Longrightarrow\ & E^{\Delta_A}_{\mathrm{frz}}\cdot\Delta^x_A = -(E^{\Delta_A}_{\mathrm{frz}})\cdot(E_A^{\Delta_A\Delta_A})^{-1}\cdot\left(E_A^{\Delta_A\mathbf{S}_A}\cdot\mathbf{S}^x_A + E_A^{\Delta_A\mathbf{h}_A}\cdot\mathbf{h}^x_A + E_A^{\Delta_A\mathrm{II}_A}\cdot\mathrm{II}^x_A + E^{\Delta_A x}_{\mathrm{xc},A}\right) \qquad \text{(B.9)}
\end{aligned}$$

We note that the symbol "II" refers to the two-electron AO integrals required for building $\mathbf{J}$ and $\mathbf{K}$, where the part for $\mathbf{K}$ is scaled by κ ($0 \le \kappa \le 1$) based on the employed density functional.

Defining $\mathbf{z}_A = -(E^{\Delta_A}_{\mathrm{frz}})\cdot(E_A^{\Delta_A\Delta_A})^{-1}$, we have

$$E^{\Delta_A}_{\mathrm{frz}}\cdot\Delta^x_A = -\mathbf{z}_A\cdot\left(E_A^{\Delta_A\mathbf{S}_A}\cdot\mathbf{S}^x_A + E_A^{\Delta_A\mathbf{h}_A}\cdot\mathbf{h}^x_A + E_A^{\Delta_A\mathrm{II}_A}\cdot\mathrm{II}^x_A + E^{\Delta_A x}_{\mathrm{xc},A}\right) \tag{B.10}$$

where $\mathbf{z}_A$ can be obtained by solving the following linear equation:

$$E_A^{\Delta_A\Delta_A}\cdot\mathbf{z}_A = -E^{\Delta_A}_{\mathrm{frz}} \tag{B.11}$$

The LHS of Eq. (B.11) contains the SCF Hessian of fragment A:

$$\begin{aligned} E_A^{\Delta_A\Delta_A} &= \frac{\partial^2 E}{\partial\Delta^{ai}\partial\Delta^{bj}} \\ &= 2(\epsilon_a-\epsilon_i)\delta_{ij}\delta_{ab} + 4(ia|jb) - 2\kappa(ij|ab) - 2\kappa(ib|ja) + 4(f_{\mathrm{xc}})_{ia,jb} \end{aligned} \tag{B.12}$$

where κ is the proportion of exact exchange in the employed density functional, $(f_{\mathrm{xc}})_{ia,jb}$ is the second functional derivative $\delta^2 E_{\mathrm{xc}}/\delta\rho(\mathbf{r})\delta\rho'(\mathbf{r})$ expressed in the MO basis. The fragment index "A" in Eq. (B.12) are omitted.

The RHS of Eq. (B.11) is the gradient of the SCF-MI energy with respect to the orbital rotations on fragment A:

$$\begin{aligned} \frac{\partial E_{\mathrm{frz}}}{\partial\Delta^{AaAi}} &= \sum_{XY}\frac{\partial E}{\partial P^{X\mu Y\nu}}\frac{\partial P^{X\mu Y\nu}}{\partial\Delta^{AaAi}} \\ &= \sum_{XY} F_{Y\nu X\mu}\left[\frac{\partial (C_o)^{X\mu}_{\ \ Xj}}{\partial\Delta^{AaAi}}(\sigma^{-1})^{XjYk}(C_o)^{\ \ Y\nu}_{Y_k} + (C_o)^{X\mu}_{\ \ Xj}(\sigma^{-1})^{XjYk}\frac{\partial (C_o)^{\ \ Y\nu}_{Y_k}}{\partial\Delta^{AaAi}}\right. \\ &\quad \left. - (C_o)^{X\mu}_{\ \ Xj}(\sigma^{-1})^{XjZl}\frac{\partial\sigma_{ZlWm}}{\partial\Delta^{AaAi}}(\sigma^{-1})^{WmYk}(C_o)^{\ \ Y\nu}_{Y_k}\right] \end{aligned} \tag{B.13}$$

Using

$$\begin{aligned} \frac{\partial (C_o)^{X\mu}_{\ \ Xj}}{\partial\Delta^{AaAi}} &= \frac{\partial}{\partial\Delta^{AaAi}}\left[C^{X\mu}_{\ \ Xp}(\delta^{Xp}_{\ \ Xi} + \Delta^{XpXq}\sigma_{XqXj} - \Delta^{XqXp}\sigma_{XqXj})\right] \\ &= C^{X\mu}_{\ \ Xp}\delta^p_{\ a}\delta^q_{\ i}\delta^X_{\ A}\sigma_{XqXj} \\ &= C^{X\mu}_{\ \ Xa}\delta^X_A\sigma_{XqXj} \end{aligned} \tag{B.14}$$

we can work out the SCF-MI gradient:

$$\frac{\partial E_{\mathrm{frz}}}{\partial\Delta^{AaAi}} = 2\sigma_{AiAj}\left[(\sigma^{-1})C_o^T F(I-PS)C_v\right]^{Aj}_{\ \ Aa} \tag{B.15}$$

Substituting Eq. (B.12) and (B.15) into Eq. (B.11), the z-vector equation becomes

$$\begin{aligned} &\left[(\epsilon_a-\epsilon_i)\delta_{ij}\delta_{ab} + 2(ia|jb) - \kappa(ij|ab) - \kappa(ib|ja) + 2(f_{\mathrm{xc}})_{ia,jb}\right](z_A)^{bj} \\ &= -\sigma_{AiAj}\left[(\sigma^{-1})C_o^T F(I-PS)C_v\right]^{Aj}_{\ \ Aa} \end{aligned} \tag{B.16}$$

which can be solved by iterative methods (e.g., the conjugate gradient method). Again, we note that the MOs on the LHS of Eq. (B.16) are on fragment A exclusively.

From z-vector to fragment response gradient

Now we turn to the terms contracted with $\mathbf{z}_A$ on the RHS of Eq. (B.10). Since all the involved AO or MO indices are on one single fragment exclusively, we omit the fragment index "A" in the following equations.

We first write down the Fock matrix as $\mathbf{F} = \mathbf{h} + \mathbf{II} \cdot \mathbf{P} + \mathbf{V}_{\text{xc}}$, and define

$$\mathbf{P}_z = \mathbf{C}_v \mathbf{z} \mathbf{C}_o^T + \mathbf{C}_o \mathbf{z}^T \mathbf{C}_v^T \tag{B.17}$$

which is a symmetric, density-like matrix. The terms on the RHS of Eq. (B.10) are derived as below:

- $\mathbf{z}_A \cdot E_A^{\Delta_A \mathbf{h}_A} \cdot \mathbf{h}_A^x$:

$$\begin{aligned} z_{ai} \cdot \frac{\partial^2 E}{\partial \Delta^{ai} \partial h_{\mu\nu}} \cdot h_{\mu\nu}^x &= z_{ai} \cdot \frac{\partial}{\partial h_{\mu\nu}} \left(\frac{\partial E}{\partial P^{\lambda\sigma}} \frac{\partial P^{\lambda\sigma}}{\partial \Delta^{ai}} \right) \cdot h_{\mu\nu}^x \\ &= z_{ai} \cdot \frac{\partial}{\partial h_{\mu\nu}} \left(F_{\lambda\sigma} [C^\lambda{}_a C^\sigma{}_i + C^\lambda{}_i C^\sigma{}_a] \right) \cdot h_{\mu\nu}^x \\ &= \delta^\mu{}_\lambda \delta^\nu{}_\sigma P_z^{\lambda\sigma} h_{\mu\nu}^x = \text{Tr}[\mathbf{P}_z \mathbf{h}^x] \end{aligned} \tag{B.18}$$

- $\mathbf{z}_A \cdot E_A^{\Delta_A \mathbf{II}_A} \cdot \mathbf{II}_A^x$:

$$\begin{aligned} z_{ai} \cdot \frac{\partial^2 E}{\partial \Delta^{ai} \partial \text{II}_{\mu\nu\lambda\sigma}} \cdot \text{II}_{\mu\nu\lambda\sigma}^x &= z_{ai} \cdot \frac{\partial}{\partial \text{II}_{\mu\nu\lambda\sigma}} \left(F_{\pi\omega} [C^\pi{}_a C^\omega{}_i + C^\pi{}_i C^\omega{}_a] \right) \cdot \text{II}_{\mu\nu\lambda\sigma}^x \\ &= P_z^{\pi\omega} \delta^\mu{}_\pi \delta^\nu{}_\omega P^{\lambda\sigma} \text{II}_{\mu\nu\lambda\sigma}^x \\ &= \text{Tr}[\mathbf{P}_z \mathbf{II}^x \mathbf{P}] \end{aligned} \tag{B.19}$$

- $\mathbf{z}_A \cdot E_{\text{xc},A}^{\Delta_A x}$ (DFT only):

$$\begin{aligned} z_{ai} \cdot \frac{\partial^2 E_{\text{xc}}}{\partial \Delta^{ai} \partial x} &= z_{ai} \cdot \frac{\partial}{\partial x} \left(\frac{\partial E_{\text{xc}}}{\partial P^{\mu\nu}} \frac{\partial P_{\mu\nu}}{\partial \Delta^{ai}} \right) \\ &= z_{ai} \cdot \frac{\partial}{\partial x} \left((V_{\text{xc}})_{\mu\nu} [C^\mu{}_a C^\nu{}_i + C^\mu{}_i C^\nu{}_a] \right) \\ &= (V_{\text{xc}}^x)_{\mu\nu} \cdot P_z^{\mu\nu} = \text{Tr}[\mathbf{P}_z \mathbf{V}_{\text{xc}}^x] \end{aligned} \tag{B.20}$$

- $\mathbf{z}_A \cdot E^{\Delta_A \mathbf{S}_A} \cdot \mathbf{S}_A^x$:

$$\begin{aligned} z_{ai} \cdot \frac{\partial^2 E}{\partial \Delta^{ai} \partial S_{\mu\nu}} \cdot S_{\mu\nu}^x &= z_{ai} \cdot \frac{\partial}{\partial S_{\mu\nu}} \left(\frac{\partial E}{\partial P^{\lambda\sigma}} \frac{\partial P^{\lambda\sigma}}{\partial \Delta^{ai}} \right) \\ &= z_{ai} \cdot \frac{\partial^2 E}{\partial P^{\lambda\sigma} \partial S_{\mu\nu}} \frac{\partial P^{\lambda\sigma}}{\partial \Delta^{ai}} \cdot S_{\mu\nu}^x + z_{ai} \cdot F_{\lambda\sigma} \frac{\partial^2 P^{\lambda\sigma}}{\partial \Delta^{ai} \partial S_{\mu\nu}} \cdot S_{\mu\nu}^x \end{aligned} \tag{B.21}$$

  The derivative of orthogonal MOs with respect to the AO overlap matrix (similar to the derivation of Eq. (B.5)):

$$\begin{aligned} \frac{\partial C^\pi{}_s}{\partial S_{\mu\nu}} &= -\frac{1}{2} C^\pi{}_r \frac{\partial \sigma_{rs}}{\partial S_{\mu\nu}} \\ &= -\frac{1}{2} C^\pi{}_r C^\lambda{}_r \delta^\mu{}_\lambda \delta^\nu{}_\sigma C^\sigma{}_s \\ &= -\frac{1}{2} (S^{-1})^{\pi\mu} C^\nu{}_s \end{aligned} \tag{B.22}$$

Based on Eq. (B.22), the first term in Eq. (B.21):

$$\begin{aligned}
z_{ai} \cdot \frac{\partial^2 E}{\partial P^{\lambda\sigma} \partial S_{\mu\nu}} \frac{\partial P^{\lambda\sigma}}{\partial \Delta^{ai}} \cdot S^x_{\mu\nu} &= z_{ai} \cdot \frac{\partial F_{\lambda\sigma}}{\partial S_{\mu\nu}} \frac{\partial P^{\lambda\sigma}}{\partial \Delta^{ai}} \cdot S^x_{\mu\nu} \\
&= P^{\lambda\sigma}_z \cdot \frac{\partial F_{\lambda\sigma}}{\partial P^{\pi\omega}} \frac{\partial P^{\pi\omega}}{\partial S_{\mu\nu}} \cdot S^x_{\mu\nu} \\
&= P^{\lambda\sigma}_z \cdot (\mathrm{II} + f_{\mathrm{xc}})_{\lambda\sigma\pi\omega} \frac{\partial P^{\pi\omega}}{\partial S_{\mu\nu}} \cdot S^x_{\mu\nu} \\
&= P^{\lambda\sigma}_z \cdot (\mathrm{II} + f_{\mathrm{xc}})_{\lambda\sigma\pi\omega} \left[-\frac{1}{2} (S^{-1})^{\pi\mu} P^{\nu\omega} - \frac{1}{2} P^{\pi\nu} (S^{-1})^{\mu\omega} \right] \cdot S^x_{\mu\nu} \\
&= -\frac{1}{2} P^{\lambda\sigma}_z (\mathrm{II} + f_{\mathrm{xc}})_{\lambda\sigma\pi\omega} \left[(S^{-1})^{\pi\mu} S^x_{\mu\nu} P^{\nu\omega} + P^{\pi\nu} S^x_{\nu\mu} (S^{-1})^{\mu\omega} \right] \\
&= -\frac{1}{2} \mathrm{Tr} \left[\mathbf{P}_z (\mathbf{II} + \boldsymbol{f}_{\mathrm{xc}}) [(\mathbf{S}^{-1}) \mathbf{S}^x \mathbf{P} + \mathbf{P} \mathbf{S}^x (\mathbf{S}^{-1})] \right]
\end{aligned} \tag{B.23}$$

and the second term:

$$\begin{aligned}
z_{ai} \cdot F_{\lambda\sigma} \frac{\partial^2 P^{\lambda\sigma}}{\partial \Delta^{ai} \partial S_{\mu\nu}} \cdot S^x_{\mu\nu} &= z_{ai} F_{\lambda\sigma} \left[-\frac{1}{2} (S^{-1})^{\lambda\mu} C^{\nu}{}_{a} C_i{}^{\sigma} - \frac{1}{2} C^{\lambda}{}_{a} C_i{}^{\nu} (S^{-1})^{\mu\sigma} \right. \\
&\quad \left. -\frac{1}{2} (S^{-1})^{\lambda\mu} C^{\nu}{}_{i} C_a{}^{\sigma} - \frac{1}{2} C^{\lambda}{}_{i} C_a{}^{\nu} (S^{-1})^{\mu\sigma} \right] \cdot S^x_{\mu\nu} \\
&= -\frac{1}{2} F_{\lambda\sigma} \cdot \left[(S^{-1})^{\lambda\mu} S^x_{\mu\nu} P^{\nu\sigma}_z + P^{\lambda\nu}_z S^x_{\nu\mu} (S^{-1})^{\mu\sigma} \right] \\
&= -\frac{1}{2} \mathrm{Tr} \left[\mathbf{F} [(\mathbf{S}^{-1}) \mathbf{S}^x \mathbf{P}_z + \mathbf{P}_z \mathbf{S}^x (\mathbf{S}^{-1})] \right]
\end{aligned} \tag{B.24}$$

Up to this point, we have derived the RHS of Eq.(B.10). Thus, fragment A's contribution to the response gradient can be expressed as

$$\begin{aligned}
E^{\Delta_A}_{\mathrm{frz}} \cdot \Delta^x_A = &- \mathrm{Tr}[\mathbf{P}_z \mathbf{h}^x] - \mathrm{Tr}[\mathbf{P}_z \mathbf{I}\mathbf{I}^x \mathbf{P}] - \mathrm{Tr}[\mathbf{P}_z \mathbf{V}^x_{\mathrm{xc}}] \\
&+ \frac{1}{2} \mathrm{Tr} \left[\mathbf{P}_z (\mathbf{II} + \boldsymbol{f}_{\mathrm{xc}}) [(\mathbf{S}^{-1}) \mathbf{S}^x \mathbf{P} + \mathbf{P} \mathbf{S}^x (\mathbf{S}^{-1})] \right] \\
&+ \frac{1}{2} \mathrm{Tr} \left[\mathbf{F} [(\mathbf{S}^{-1}) \mathbf{S}^x \mathbf{P}_z + \mathbf{P}_z \mathbf{S}^x (\mathbf{S}^{-1})] \right]
\end{aligned} \tag{B.25}$$

B.1.3 Evaluation of matrix-vector products using finite difference

z-vector contracted with the implicit first derivative of $\mathbf{V}_{\mathbf{xc}}$

Eq. (B.23) can be used to evaluate the first term of Eq. (B.21) when the analytical form of the second functional derivative ($\boldsymbol{f}_{\mathrm{xc}}$, also known as the implicit first derivative of $\mathbf{V}_{\mathrm{xc}}$) of the employed density functional is available. For ωB97X-V that is used in this work, we can compute the entire contribution from the XC part ($\mathbf{z}_A \cdot E^{\Delta_A \mathbf{S}_A}_{\mathrm{xc}} \cdot \mathbf{S}^x_A$) in a different way using finite difference.

$$z_{ai} \cdot \frac{\partial^2 E_{\mathrm{xc}}}{\partial \Delta^{ai} \partial S_{\mu\nu}} \cdot S^x_{\mu\nu} = z_{ai} \cdot \frac{\partial^2 E_{\mathrm{xc}}}{\partial \Delta^{ai} \partial P^{\pi\omega}} \frac{\partial P^{\pi\omega}}{\partial S_{\mu\nu}} \cdot S^x_{\mu\nu} = \left(z_{ai} \cdot \frac{\partial (V_{\mathrm{xc}})_{\pi\omega}}{\partial \Delta^{ai}} \right) \left(\frac{\partial P^{\pi\omega}}{\partial S_{\mu\nu}} \cdot S^x_{\mu\nu} \right) \tag{B.26}$$

Using the finite-difference matrix-vector product technique,

$$z_{ai} \cdot \frac{\partial (V_{\mathrm{xc}})_{\pi\omega}}{\partial \Delta_{ai}} = \frac{(V_{\mathrm{xc}})_{\pi\omega}[\mathbf{P}_{+\delta z}] - (V_{\mathrm{xc}})_{\pi\omega}[\mathbf{P}_{-\delta z}]}{2\delta} \tag{B.27}$$

where δ is the step size for finite-difference calculations (in practice we use $\delta = 10^{-4}$). $\mathbf{P}_{+\delta z}$ is the density matrix constructed upon (occupied) MOs updated as follows (a step forward of size $\delta \mathbf{z}$):

$$\mathbf{C}' = \mathbf{C} \exp \begin{pmatrix} 0 & -\delta \mathbf{z}^T \\ \delta \mathbf{z} & 0 \end{pmatrix} \tag{B.28}$$

Based on Eq. (B.27), the evaluation of one matrix-vector product only requires constructing the $\mathbf{V}_{\mathrm{xc}}$ matrix twice (with $\mathbf{P}_{+\delta z}$ and $\mathbf{P}_{-\delta z}$). And the term left has been worked out before

$$\frac{\partial P^{\pi\omega}}{\partial S_{\mu\nu}} \cdot S^x_{\mu\nu} = -\frac{1}{2} \left[(S^{-1})^{\pi\mu} S^x_{\mu\nu} P^{\nu\omega} + P^{\pi\mu} S^x_{\mu\nu} (S^{-1})^{\nu\omega} \right] \tag{B.29}$$

We note that if we multiply the RHS of Eq. (B.27) by

$$\frac{\partial P^{\pi\omega}}{\partial \Delta^{bj}} = C^{\pi}_{\ b} C^{\omega}_{\ j} + C^{\pi}_{\ j} C^{\omega}_{\ b} \tag{B.30}$$

we can get the result for $\mathbf{z}_A \cdot E^{\Delta_A \Delta_A}$, which is necessary for iteratively solving the z-vector equation.

z-vector contracted with the explicit first derivative of $\mathbf{V}_{\mathrm{xc}}$

The use of Eq. (B.20) for the evaluation of $\mathbf{z}_A \cdot E^{\Delta_A x}_{\mathrm{xc},A}$ requires the analytical form of $\mathbf{V}^x_{\mathrm{xc}}$, which is the explicit first derivative of the XC matrix. If that is not available for the employed functional, we can also compute this quantity using finite difference. Analogous to Eq. (B.27), we have

$$z_{ai} \cdot \frac{\partial^2 E_{\mathrm{xc}}}{\partial \Delta_{ai} \partial x} = \frac{E^x_{\mathrm{xc}}[\mathbf{P}_{+\delta z}] - E^x_{\mathrm{xc}}[\mathbf{P}_{-\delta z}]}{2\delta} \tag{B.31}$$

where $\mathbf{P}_{+\delta z}$ has the same definition as above. E^x_{xc} is the explicit first derivative of the XC energy, which is related to the change of integration quadrature with respect to the displacement of nuclei. Based on Eq. (B.31), it only requires computing E^x_{xc} twice to obtain $\mathbf{z}_A \cdot E^{\Delta_A x}_{\mathrm{xc},A}$.

B.2 Additional Results

Table B.1: Adiabatic EDA results for the water dimer in linear, bifurcated and aligned configurations computed at the B3LYP/def2-TZVPPD level of theory. Other details are the same as in Table 3.1 (in Chap. 3)

	linear			bifurcated			aligned		
	FRZ	POL	FULL	FRZ	POL	FULL	FRZ	POL	FULL
Adiabatic ΔE	-11.18	-2.43	-5.43	-8.15	-1.19	-1.23	-5.63	-0.79	-0.64
E_{bind}	-11.18	-13.61	-19.03	-8.15	-9.34	-10.57	-5.63	-6.42	-7.06
$R(\text{O}_2 \cdot\cdot \text{H}_d)$	2.28	2.18	1.96	2.74	2.66	2.57	2.91	2.83	2.76
$R(\text{O}_1 \cdot\cdot \text{O}_2)$	3.24	3.13	2.92	3.23	3.15	3.06	3.41	3.32	3.25
$R(\text{O}_1\text{–H}_d)$	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96
α angle (°)	172.66	172.76	173.04	–	–	–	–	–	–
β angle (°)	132.78	132.69	126.36	–	–	–	–	–	–
ω_1	96.86	106.13	122.51	152.26i	182.34i	231.82i	178.96i	212.76i	241.12i
ω_2	98.16	116.08	149.78	62.34	71.34	73.56	139.53i	151.04i	164.91i
ω_9	3809.58	3803.23	3692.89	3810.99	3810.55	3808.50	3811.64	3811.45	3810.56
ω_{10}	3818.05	3809.78	3804.59	3821.20	3821.71	3817.52	3821.82	3822.92	3822.14
ω_{11}	3910.15	3905.30	3885.02	3908.17	3904.93	3895.86	3910.81	3909.10	3905.88
ω_{12}	3914.97	3908.64	3904.31	3912.17	3911.47	3910.26	3911.55	3910.84	3910.17
$\omega_{10} - \omega_9$	8.47	6.55	111.70	10.21	11.16	9.02	10.18	11.47	11.58
$\omega_{12} - \omega_{11}$	4.82	3.34	19.29	4.00	6.54	14.40	0.74	1.74	4.29

Note: compared to the results in Chap. B computed with the ωB97X-V functional, most of the original conclusions still hold. For instance, we still see that the linear configuration is already preferred at the FRZ level, while the red shift of ω_9 (symmetric O–H stretch of the H-donor water) turns out to be an effect of CT. Nevertheless, quantitative differences do exist between the B3LYP and ωB97X-V results. It is clear that B3LYP underbinds the water dimer relative to ωB97X-V, and the resulting intermolecular separations for bifurcated and aligned configurations are significantly larger here. This is most likely because (i) B3LYP fails to properly describe long-range dispersion and (ii) B3LYP slightly underestimates the dipole moment of water molecule compared to that given by ωB97X-V.

Table B.2: Adiabatic EDA results for the water-Cl^- complex computed at the B3LYP/def2-TZVPPD level of theory. Other details are the same as in Table 3.3 in Chap. 3.

	FRZ (C_s)	POL (C_s)	FULL (C_s)	FRZ (C_{2v})	POL (C_{2v})	FULL (C_{2v})
Adiabatic ΔE	-36.05	-9.03	-14.43	-35.97	-8.23	-7.36
E_{bind}	-36.05	-45.07	-59.50	-35.97	-44.20	-51.56
R (Cl··O)	3.47	3.38	3.13	3.43	3.33	3.17
R (Cl··H_d)	2.72	2.50	2.15	2.91	2.80	2.63
R (O··H_d)	0.96	0.97	0.99	0.96	0.96	0.97
∠Cl–H_d–O	134.37	151.25	168.57	115.44	115.48	115.70
∠H–O–H	100.39	100.90	101.78	99.77	98.73	96.69
ω_1	71.26	135.18	190.65	$98.03i$	$116.43i$	$328.62i$
ω_5	3812.42	3748.51	3311.63	3819.60	3807.36	3744.96
ω_6	3877.55	3874.99	3869.45	3871.85	3847.54	3753.34
split ($\omega_6 - \omega_5$)	65.13	126.48	557.82	52.25	40.18	8.38

Note: compared to the results by ωB97X-V, there are many similarities such as the magnitude of the split between two O–H stretches on the polarized and fully relaxed PESs. The major difference is that there appears an additional stationary point on the frozen PES wherein Cl^- deviates from the bisector of ∠H–O–H. It is marginally more advantageous (by only 0.08 kJ/mol) than another stationary structure (C_{2v}) in terms of stabilization energy. Therefore, the energy minimum on the frozen PES has a broken-symmetry geometry, which might be related to the lack of long-range dispersion in B3LYP (see Table B.3). Besides that, B3LYP also underbinds the water-Cl^- complex for the same reasons as in the water dimer scenario.

Table B.3: Adiabatic EDA results for the water-Cl^- complex computed at the B3LYP-D3/def2-TZVPPD level of theory. Other details are the same as in Table 3.3 in Chap. 3.

	FRZ (C_{2v})	POL (C_{2v})	FULL (C_{2v})	FULL (C_s)
Adiabatic ΔE	-40.25	-8.46	-7.29	-13.63
E_{bind}	-40.25	-48.71	-56.00	-62.34
R (Cl··O)	3.39	3.31	3.20	3.13
R (Cl··H_d)	2.86	2.78	2.66	2.16
R (O··H_d)	0.96	0.96	0.97	0.99
∠Cl–H_d–O	115.30	115.41	115.67	167.31
∠H–O–H	99.61	98.71	96.99	101.51
ω_1	149.62	135.83	$284.49i$	191.94
ω_5	3821.39	3808.56	3748.17	3320.91
ω_6	3873.10	3849.40	3761.07	3868.97
split ($\omega_6 - \omega_5$)	51.71	40.84	12.90	548.06

Note: adding a D3 correction to B3LYP, the C_{2v} symmetry is restored for the energy minimum on the frozen PES. The resulting energetics is also more similar to that given by ωB97X-V. However, on the polarized PES, the minimum-energy structure retains C_{2v} symmetry (verified by frequency calculation), which is different from the previous results given by ωB97X-V or B3LYP. In our previous calculations, we have noticed that two stationary points on the polarized PES are of very close stabilization energies, and here their energetic order is overturned due to the change of functional.

Table B.4: Basis set dependence of the N–B bond length (Å) and the N–B–H angle (°) at the optimized structures of NH_3-BH_3 on the frozen, polarized and fully relaxed PESs with three triple-ζ basis sets: def2-TZVPP, def2-TZVPPD, aug-cc-pVTZ. Two different models ("ALMO" and "FERF") are employed to construct the polarized surface.

Basis set	N–B distance (Å)			
	FRZ	POL (FERF)	POL (ALMO)	FULL
def2-TZVPP	2.64	2.38	2.35	1.66
def2-TZVPPD	2.76	2.36	2.10	1.66
aug-cc-pVTZ	2.80	2.41	1.77	1.66
	The N–B–H angle (°)			
	FRZ	POL (FERF)	POL (ALMO)	FULL
def2-TZVPP	92.63	94.71	95.05	104.97
def2-TZVPPD	92.08	95.08	98.38	104.97
aug-cc-pVTZ	91.95	94.60	104.17	104.97

Note: according to the results above, the optimal values for R(N–B) and $\angle$N–B–H on the fully relaxed PES are insensitive to the choice of basis set, while their values on the ALMO-based polarized PES vary significantly with the increase of basis size. The use of the FERF model largely reduces the basis set sensitivity of the optimal structural parameters on the polarized surface. The basis set effect on the optimal structure on the frozen PES, on the other hand, has not been discussed elsewhere. Here we see that the optimal N–B distance on the frozen PES increases with the size of the employed basis set while $\angle$N–B–H decreases, which might be related to the enhancement of Pauli repulsion when fragment electron density becomes more diffuse. Nevertheless, the changes are smaller than those on the original ALMO-based polarized PES, and the qualitative interpretation of the results computed on the frozen surface is *not* affected.

Appendix C

Additional Materials for the Discussion of CT

C.1 Nuclear Gradient for the CDFT(Becke) Approach

In the derivation presented below, we use μ, ν,... for AO basis indices, i, j,... for occupied MO indices, a, b,... for virtual MO indices, and p, q,... for generic (occupied or virtual) MO indices. Tensorial notations are used following the same convention as in our previous papers (e.g. Ref. 116).

The energy functional given by Eq. (4.12) can be rewritten using the 1PDM:

$$E[\mathbf{P}] = E_0[\mathbf{P}] + \sum_A \lambda_A \,\mathrm{Tr}[(\mathbf{P} - \mathbf{P}_0)\mathbf{W}_A] \tag{C.1}$$

where E_0 is the standard KS energy, $\mathbf{W}_A$ is the weight matrix as defined in Eq. (4.11), and $\mathbf{P}_0$ is the promolecule density matrix that has a fragment-block-diagonal structure:

$$(P_0)^{A\mu A\nu} = \sum_i (C_o)^{A\mu}_{\ \ Ai} (C_o^T)_{Ai}^{\ \ A\nu} \tag{C.2}$$

Differentiating both sides of Eq. (C.1) with respect to nuclear positions (x), we get

$$\begin{aligned} E^x = E_0^x &+ \sum_A \lambda_A \,\mathrm{Tr}[(\mathbf{P} - \mathbf{P}_0)\mathbf{W}_A^x] \\ &- \sum_{A=1}^{N_{\mathrm{frgm}}} \left(\frac{\partial \mathbf{P}_0}{\partial \mathbf{S}_A} \cdot \mathbf{S}_A^x + \frac{\partial \mathbf{P}_0}{\partial \boldsymbol{\theta}_A} \cdot \boldsymbol{\theta}_A^x \right) \cdot \mathbf{W}_{\mathrm{sum}} \end{aligned} \tag{C.3}$$

where $\mathbf{S}_A$ and $\boldsymbol{\theta}_A$ refer to the AO overlap matrix and the orbital rotation matrix ($n_v \times n_o$) for fragment $\mathbf{A}$, respectively, and

$$\mathbf{W}_{\mathrm{sum}} = \sum_A \lambda_A \mathbf{W}_A \tag{C.4}$$

The first term on the RHS of Eq. (C.3) is the standard KS-DFT gradient, which has been well-documented in literature,[127] and the second term represents the change in the CDFT weight matrices with respect to the nuclear displacement, which can be obtained by modifying the routines available in Q-Chem that evaluate the XC contribution to the gradient, as the second term on the RHS of Eq. (C.1) is essentially an local density approximation (LDA)-like functional. The third term, which reflects the response of the promolecule density, is more challenging to evaluate. Starting from Eq. (C.2), we first parameterize ${C^{A\mu}}_{Ai}$ with $\mathbf{S}_A$ and $\boldsymbol{\theta}_A$:

$$\begin{aligned} {C^{A\mu}}_{Ai} = {C^{A\mu}}_{As}({S_A^{-\frac{1}{2}}})^{As}{}_{Ar}\left[{\delta^{Ar}}_{Ai} + (\theta_A)^{ArAs}\delta_{AsAi}\right. \\ \left. -(\theta_A)^{AsAr}\delta_{AsAi} + \mathcal{O}(\boldsymbol{\theta}_A^2)\right] \end{aligned} \tag{C.5}$$

Therefore, we have

$$\begin{aligned} &\sum_{A=1}^{N_{\text{frgm}}} (W_{\text{sum}})_{A\mu A\nu}\frac{\partial P_0^{A\mu A\nu}}{\partial S_{A\lambda A\sigma}}S^x_{A\lambda A\sigma} \\ &= \sum_{A=1}^{N_{\text{frgm}}} -\frac{1}{2}\left({C^{A\mu}}_{As}{C^{A\lambda}}_{As}S^x_{A\lambda A\sigma}{C^{A\sigma}}_{Ai}{C^{A\nu}}_{Ai} + \text{c.c.}\right)(W_{\text{sum}})_{A\mu A\nu} \\ &= \sum_{A=1}^{N_{\text{frgm}}} -\frac{1}{2}\left[(\mathbf{S}_A^{-1})\mathbf{S}_A^x\mathbf{P}_A + \mathbf{P}_A\mathbf{S}_A^x(\mathbf{S}_A^{-1})\right]\cdot\mathbf{W}_{\text{sum},A} \end{aligned} \tag{C.6}$$

and

$$\begin{aligned} &\sum_{A=1}^{N_{\text{frgm}}} (W_{\text{sum}})_{A\mu A\nu}\frac{\partial P_0^{A\mu A\nu}}{\partial \theta^{AaAi}}(\theta^x)^{AaAi} \\ &= \sum_{A=1}^{N_{\text{frgm}}} (W_{\text{sum}})_{A\mu A\nu}({C^{A\mu}}_{Aa}{C^{A\nu}}_{Ai} + {C^{A\mu}}_{Ai}{C^{A\nu}}_{Aa})(\theta^x)^{AaAi} \\ &= 2\sum_{A=1}^{N_{\text{frgm}}} (W_{\text{sum}})_{AaAi}(\theta^x)^{AaAi} \\ &= 2\sum_{A=1}^{N_{\text{frgm}}} \mathbf{W}^{(\text{vo})}_{\text{sum},A}\cdot\boldsymbol{\theta}^x_A \end{aligned} \tag{C.7}$$

The RHS of Eq. (C.7) can be computed using the z-vector approach,[302] i.e., based on the stationary condition of SCF calculations for isolated fragments, we have

$$\begin{aligned} \mathbf{W}^{(\text{vo})}_{\text{sum},A}\cdot\boldsymbol{\theta}^x_A = \mathbf{z}_A\cdot\Big(E_A^{\boldsymbol{\theta}_A\mathbf{h}_A}\cdot\mathbf{h}^x_A + E_A^{\boldsymbol{\theta}_A\mathbf{II}_A}\cdot\mathbf{II}^x_A \\ + E_A^{\boldsymbol{\theta}_A\mathbf{S}_A}\cdot\mathbf{S}^x_A + E^x_{\text{xc},A}\Big) \end{aligned} \tag{C.8}$$

where $\mathbf{h}_A$ is fragment A's core Hamiltonian, and $\mathbf{II}_A$ refers to its two-electron AO integrals. The z-vector, $\mathbf{z}_A$, is the solution to the following linear equation:

$$E_A^{\boldsymbol{\theta}_A\boldsymbol{\theta}_A}\mathbf{z}_A = \mathbf{W}_{\mathrm{sum},A}^{(\mathrm{vo})} \tag{C.9}$$

where $E_A^{\boldsymbol{\theta}_A\boldsymbol{\theta}_A}$ is the orbital Hessian for the KS energy of fragment A. The contraction of the z-vector with the quantity in the parenthesis in Eq. (C.8) involves rather complicated equations, and we refer the reader to the ESI of our previous paper (Ref. 119) for the full details.

We note that the nuclear gradient derived above should have been comprised in the gradient for the CDFT configuration interaction (CI) theory reported by Kaduk et al. [549] Nonetheless, for the special case where only one single determinant is involved, the form given above is much easier to understand and convenient for implementation purpose.

C.2 Additional Results

Table C.1: Adiabatic EDA results for the formic acid dimer (of C_{2h} symmetry) computed with ωB97X-V/def2-TZVPPD. E_{bind} (in kJ/mol) refers to the energy lowering relative to twice of the isolated monomer energy; r_{OH} and $r_{\text{O}\cdot\text{H}}$ (in Å) refer to the bond length of the proton-donating O–H and the intermolecular O··H distance, respectively; ∠O–H··O is reported in degrees and ν_{OH} (the OH stretch frequency) is in cm^{-1}. CDFT gives the identical results as full SCF for this system.

	Monomer	FRZ	POL (ALMO)	POL (FERF)	Full SCF
E_{bind}		-33.04	-46.60	-45.46	-70.01
r_{OH}	0.968	0.970	0.974	0.973	0.998
$r_{\text{O}\cdot\text{H}}$		2.197	2.001	2.016	1.670
∠O–H··O		173.09	173.35	173.33	178.61
ν_{OH}	3789	3791	3721	3731	3253

Table C.2: Adiabatic EDA results for the NH_3-BH_3 and BH_3-CO complexes computed with ωB97X-V/def2-TZVPP. Two definitions of the "CT-free" state, ALMO and CDFT (Becke), are compared. E_{bind} is the energy lowering (in kJ/mol) relative to the sum of isolated monomer energies. The reported distances are in Å and angles in degrees (°).

	NH_3-BH_3			
	FRZ	ALMO	CDFT	Full
E_{bind}	-19.55	-27.95	-73.77	-136.80
$R(\text{N}\cdot\cdot\text{B})$	2.64	2.35	1.72	1.66
∠N–B–H	92.62	95.05	102.63	104.98
	BH_3-CO			
	FRZ	ALMO	CDFT	Full
E_{bind}	-5.71	-6.80	-115.43	-116.00
$R(\text{B}\cdot\cdot\text{C})$	3.09	2.98	1.53	1.54
R(C–O)	1.125	1.125	1.127	1.126
∠C–B–H	90.52	90.78	103.99	103.95

Table C.3: Equilibrium CT stabilization energies (in kJ/mol) for the water-M^{n+} (M^{n+}= Li^+, Na^+, K^+, Mg^{2+} and Ca^{2+}) complexes evaluated using AO-based ALMO, FERF, CDFT(Becke) and CDFT(FBH) methods. All the calculations are performed at the ωB97X-D3/def2-QZVPPD level of theory.

	ALMO	FERF	CDFT(Becke)	CDFT(FBH)
Li^+	-10.48	-14.14	-16.46	-5.79
Na^+	-1.77	-2.36	-6.31	-3.43
K^+	-4.12	-4.56	-4.35	-4.42
Mg^{2+}	-11.99	-18.58	-60.03	-33.44
Ca^{2+}	-37.22	-40.77	-40.53	-38.92

Appendix D

Additional Results for the Assessment of AMOEBA

Table D.1: Total interaction energy and energy components (in kcal/mol) given by SAPT(DFT), ALMO-EDA, and the AMOEBA force field for the "linear" and "bifurcated" configurations of the water dimer. The SAPT(DFT) results are provided by *J. Phys. Chem. A* **2016**, 120, 4550–4559, while the results of AMOEBA and ALMO-EDA are generated with the same protocol as in the main paper. The two optimized geometries of the water dimer are taken from *J. Chem. Phys.* **2002**, 116, 690-701.

	SAPT(DFT)		ALMO-EDA		AMOEBA	
"Linear"	ELST	-8.01	CLS ELEC	-8.54	PERM ELEC	-5.91
	EXCH	7.87	PAULI	8.57	VDW (rep)	3.99
	DISP	-2.41	DISP	-1.91	VDW (attr)	-1.80
	IND	-2.29	POL	-1.09	POL	-1.20
			CT	-1.99		
	TOTAL	-4.83		-4.96		-4.92
	SAPT(DFT)		ALMO-EDA		AMOEBA	
"Bifurcated"	ELST	-4.33	CLS ELEC	-4.41	PERM ELEC	-3.12
	EXCH	3.15	PAULI	3.21	VDW (rep)	1.21
	DISP	-1.55	DISP	-1.38	VDW (attr)	-1.01
	IND	-0.57	POL	-0.42	POL	-0.40
			CT	-0.30		
	TOTAL	-3.31		-3.39		-3.34

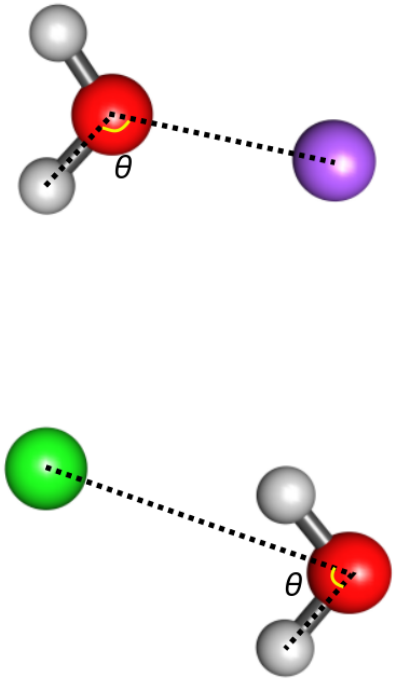

Figure D.1: Illustration of the rigid angular scans performed for water-cation (top) and water-anion (bottom) complexes. θ is the angle being modified.

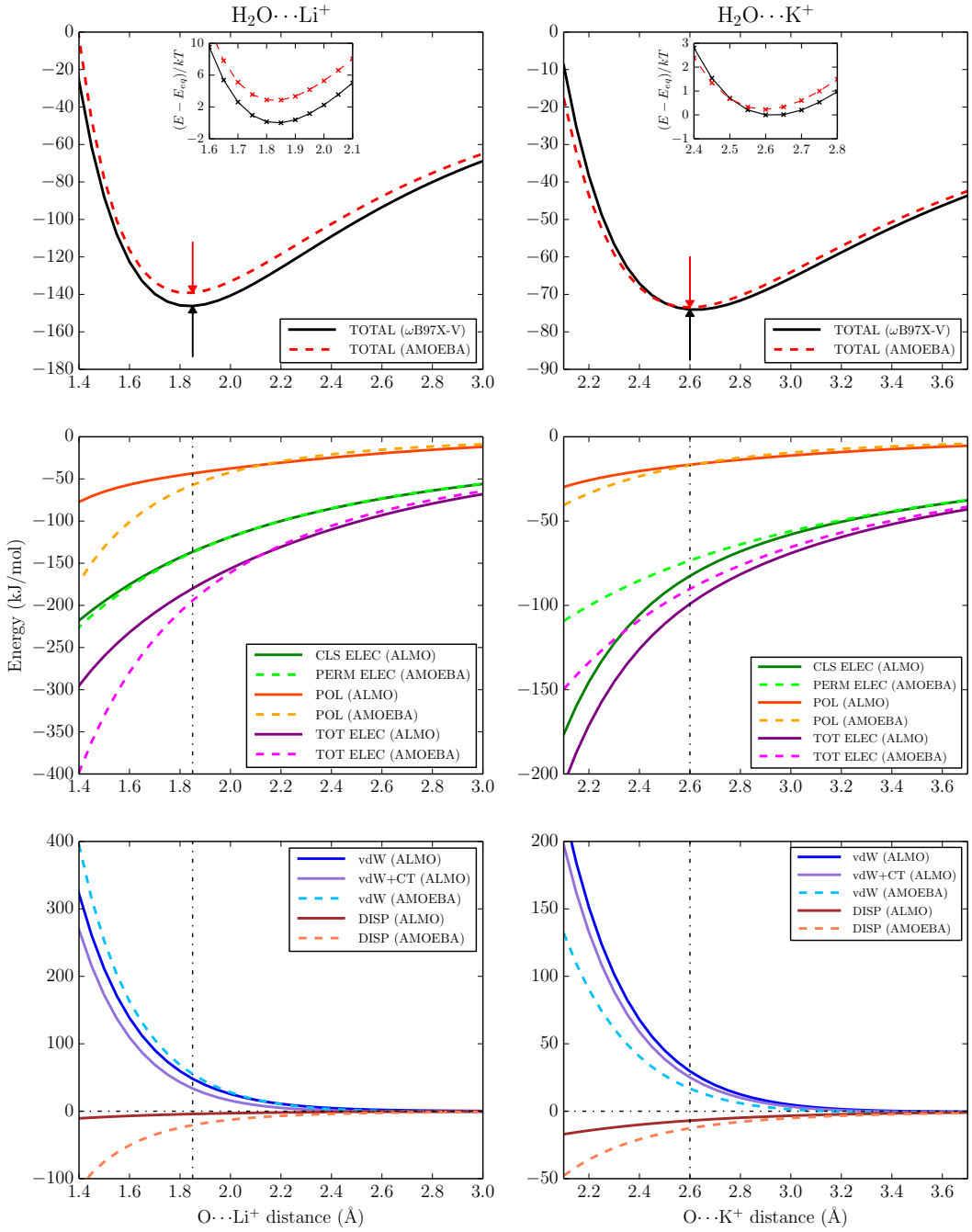

Figure D.2: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid dissociation of $H_2O\cdots Li^+$ (left) and $H_2O\cdots K^+$ (right) complexes. Top: total interaction energies; middle: permanent and induced electrostatic interactions; bottom: vdW interactions. The inset plots in the two top panels show the zoomed-in near-equilibrium region in the units of kT, and the arrows indicate the location of energy minima for QM and AMOEBA interactions, while the dash-dotted lines in the lower four panels indicate the position of QM minimum.

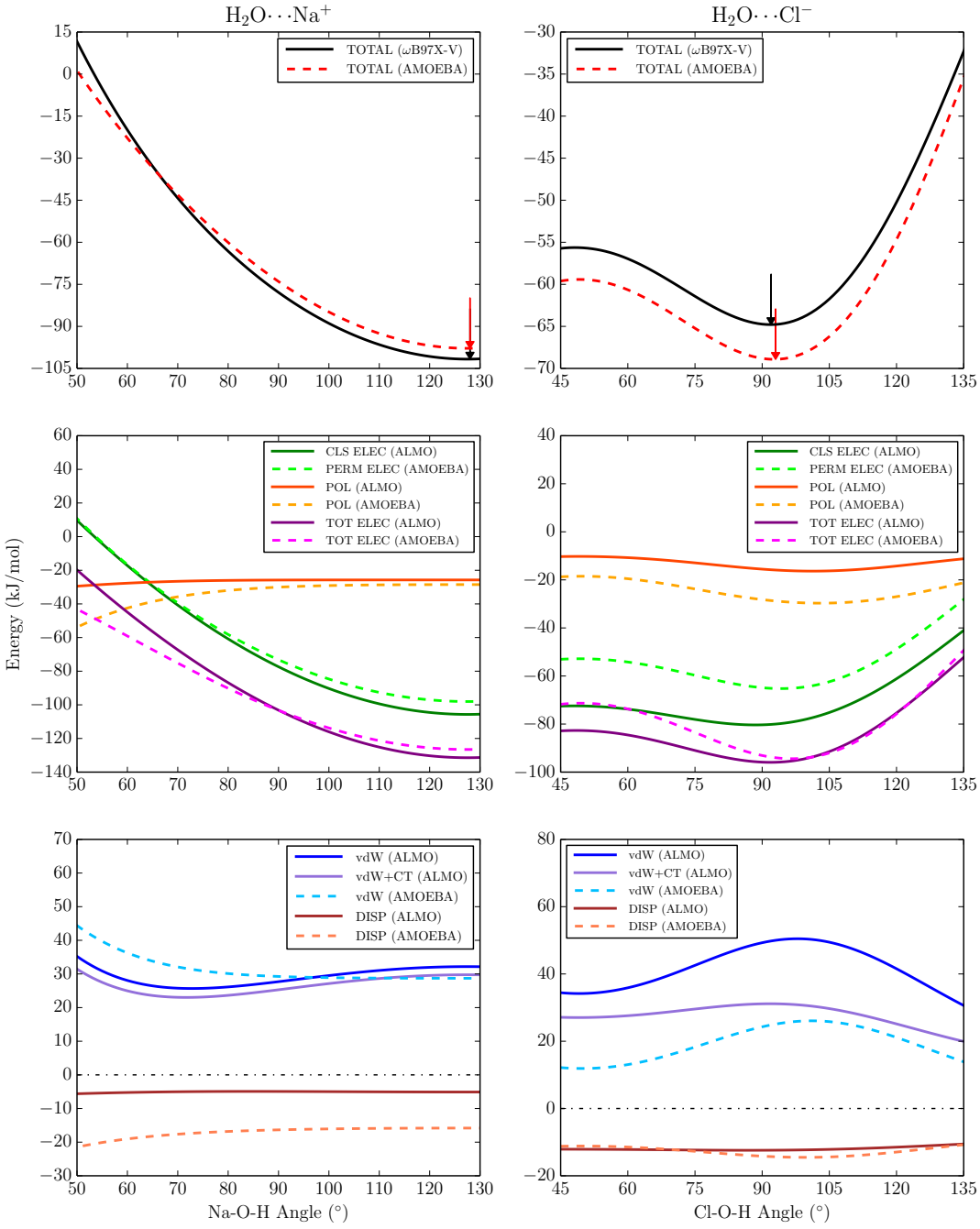

Figure D.3: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid angular scan for $H_2O\cdots Na^+$ and $H_2O\cdots Cl^-$ complexes. The angles being modified during the scan are indicated by Figure D.1. The arrows indicate the energy minima on QM and AMOEBA total interaction energy curves.

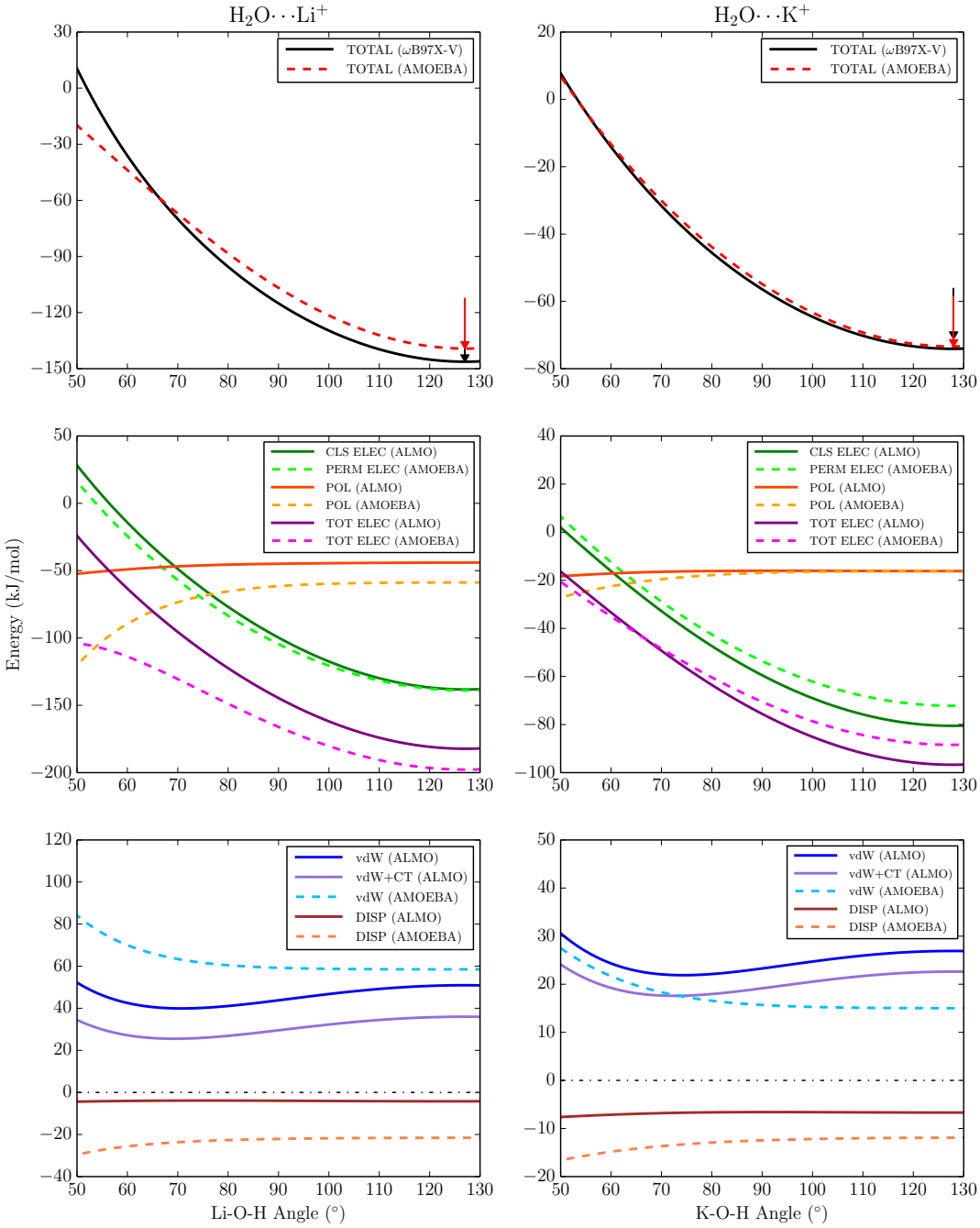

Figure D.4: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid angular scan for $H_2O\cdots Li^+$ and $H_2O\cdots K^+$ complexes. The angles being modified during the scan are identically defined as in the $H_2O\cdots Na^+$ case. The arrows indicate the energy minima on QM and AMOEBA total interaction energy curves.

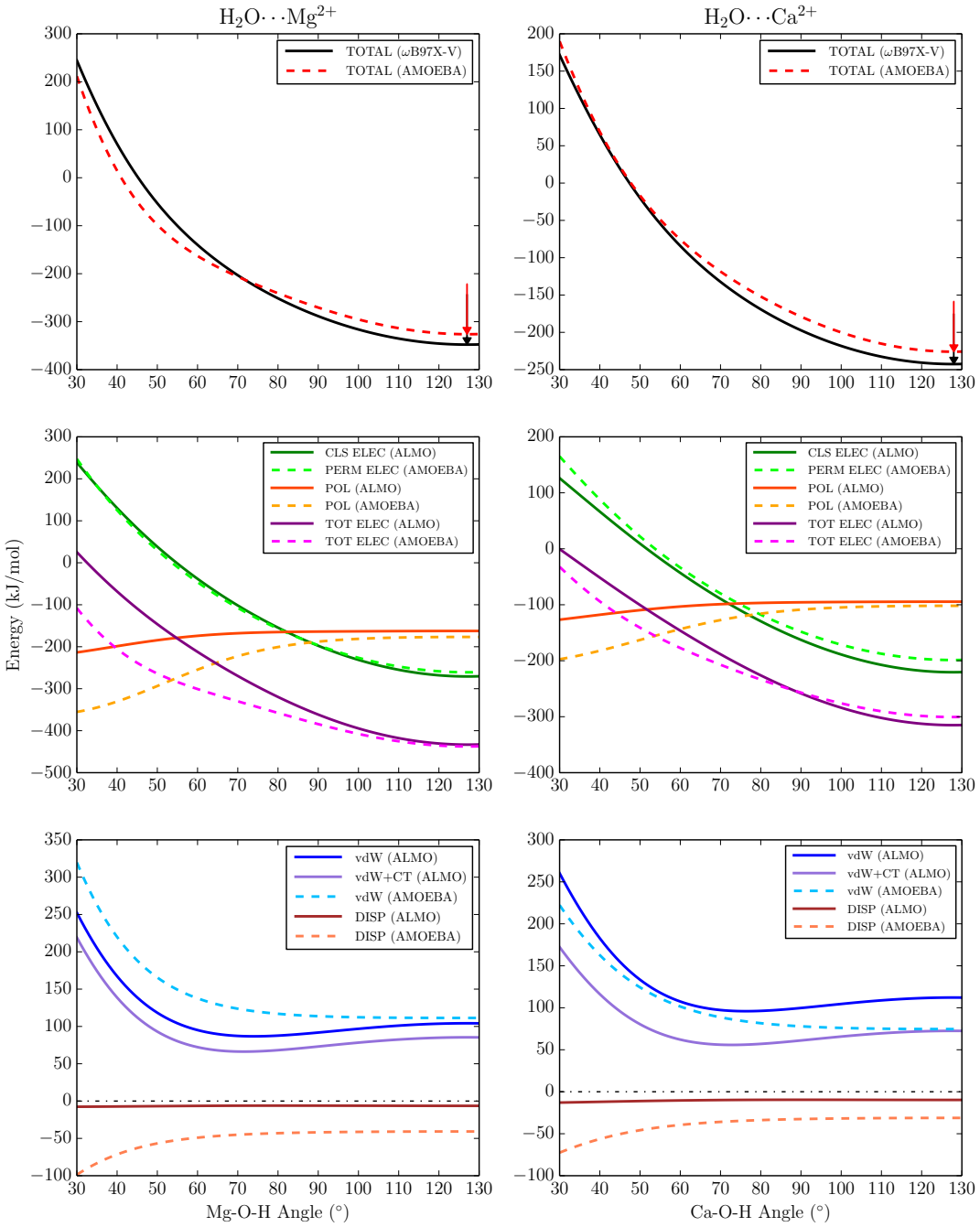

Figure D.5: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid angular scan for $H_2O\cdots Mg^{2+}$ and $H_2O\cdots Ca^{2+}$ complexes. The angles being modified during the scan are identically defined as in the $H_2O\cdots Na^{+}$ case. The arrows indicate the energy minima on QM and AMOEBA total interaction energy curves.

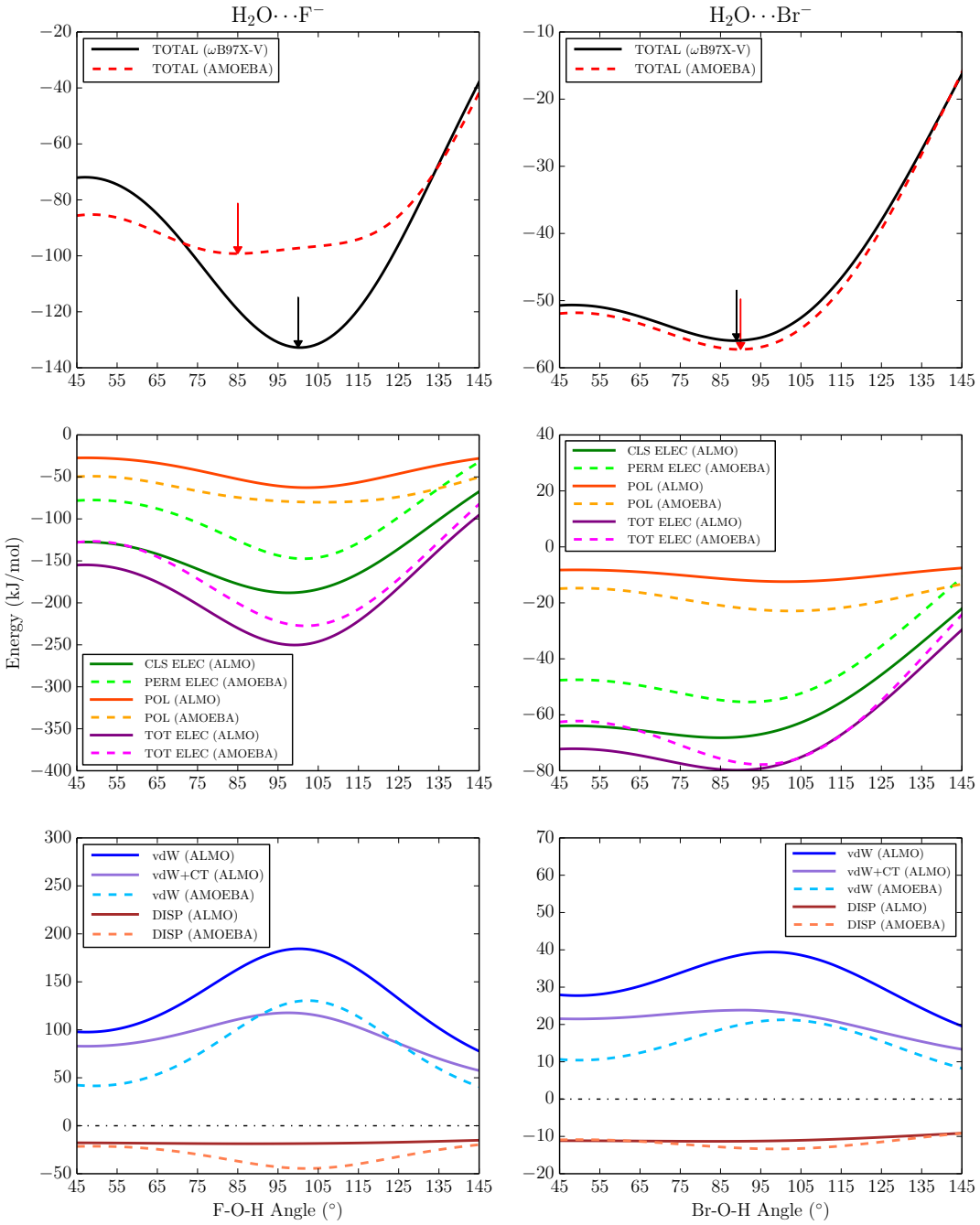

Figure D.6: The total interaction energy and its breakdowns (in kJ/mol) upon a rigid angular scan for $H_2O\cdots F^-$ and $H_2O\cdots Br^-$ complexes. The angles being modified during the scan are identically defined as in the $H_2O\cdots Cl^-$ case. The arrows indicate the energy minima on QM and AMOEBA total interaction energy curves.

Appendix E

Additional Results for QM/AMOEBA

E.1 Comparison between Polarization Energies given by ALMO and FERF

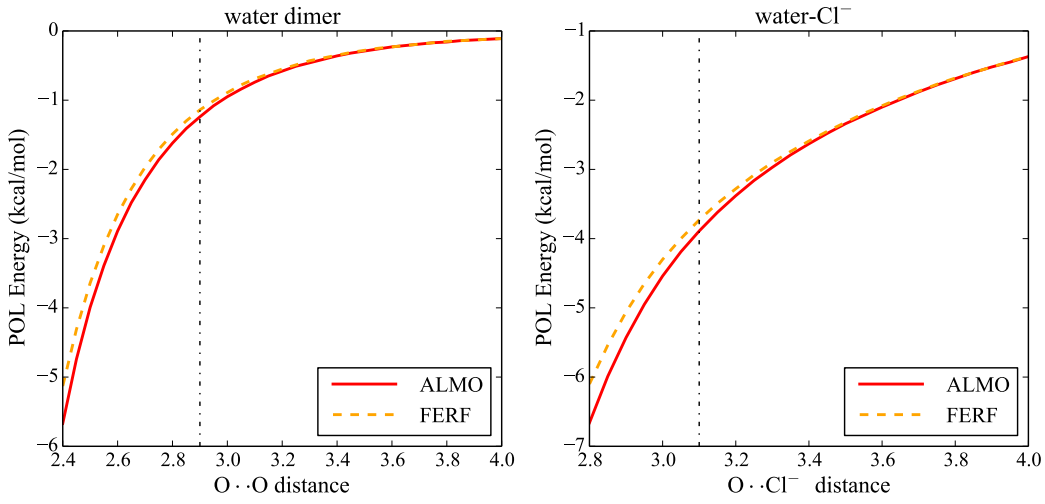

Figure E.1: Polarization energies (in kcal/mol) for the rigid dissociation of the water dimer (left) and the water-Cl^- complex (right) evaluated by the ALMO and FERF models at the ωB97X-V/def2-TZVPPD level of theory. The vertical dashed lines indicate the equilibrium distances.

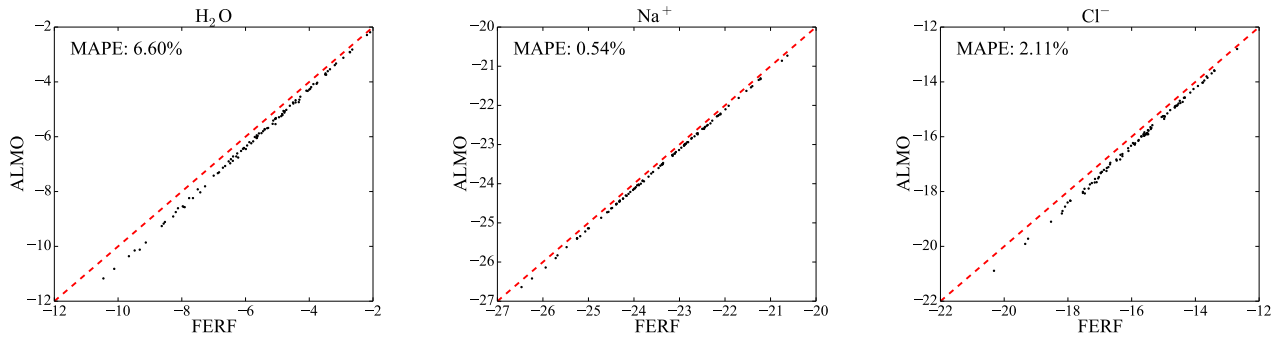

Figure E.2: Plots of full QM polarization energies (in kcal/mol) evaluated with the ALMO model (y axis) against those evaluated with the FERF model (x axis) for three solutes (H_2O, Na^+, Cl^-) interacting with the water molecules in their first solvation shells. 100 snapshots are calculated for each solute. For each point, its distance from the diagonal indicate the difference between the values given by these two models. The mean absolute percentage error (MAPE) for each system is also provided (using FERF-derived polarization energies as references).

E.2 Basis Set Dependence of Full QM and QM/AMOEBA Results

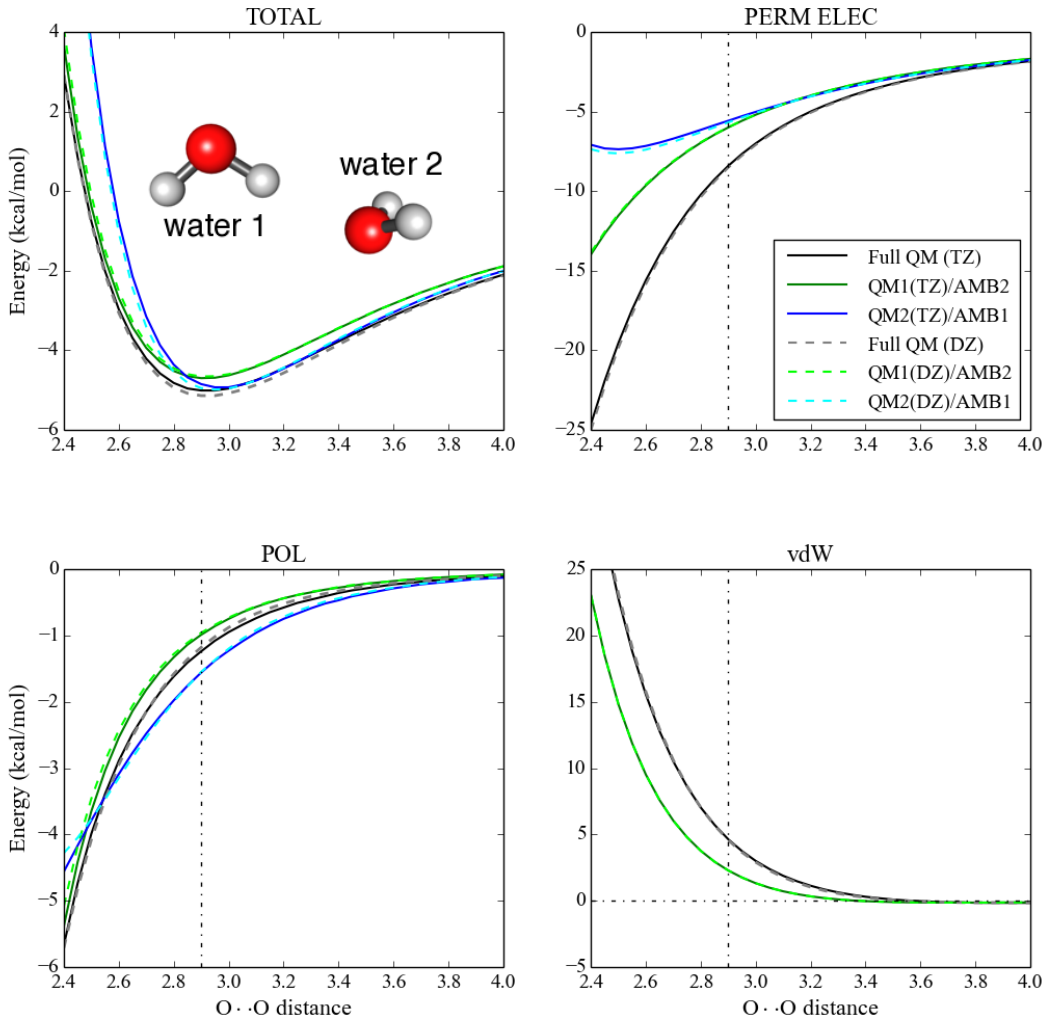

Figure E.3: Total interaction energies and their components (in kcal/mol) evaluated with full QM and QM/AMOEBA for the water dimer as functions of the O··O distance. The ωB97X-V functional and two basis sets, def2-SVPD (which is denoted as "DZ") and def2-TZVPPD ("TZ"), are employed for the QM and QM/AMOEBA calculations. The vdW interactions in QM/AMOEBA calculations are evaluated through the fully empirical 14-7 potential so that they have no basis set dependence.

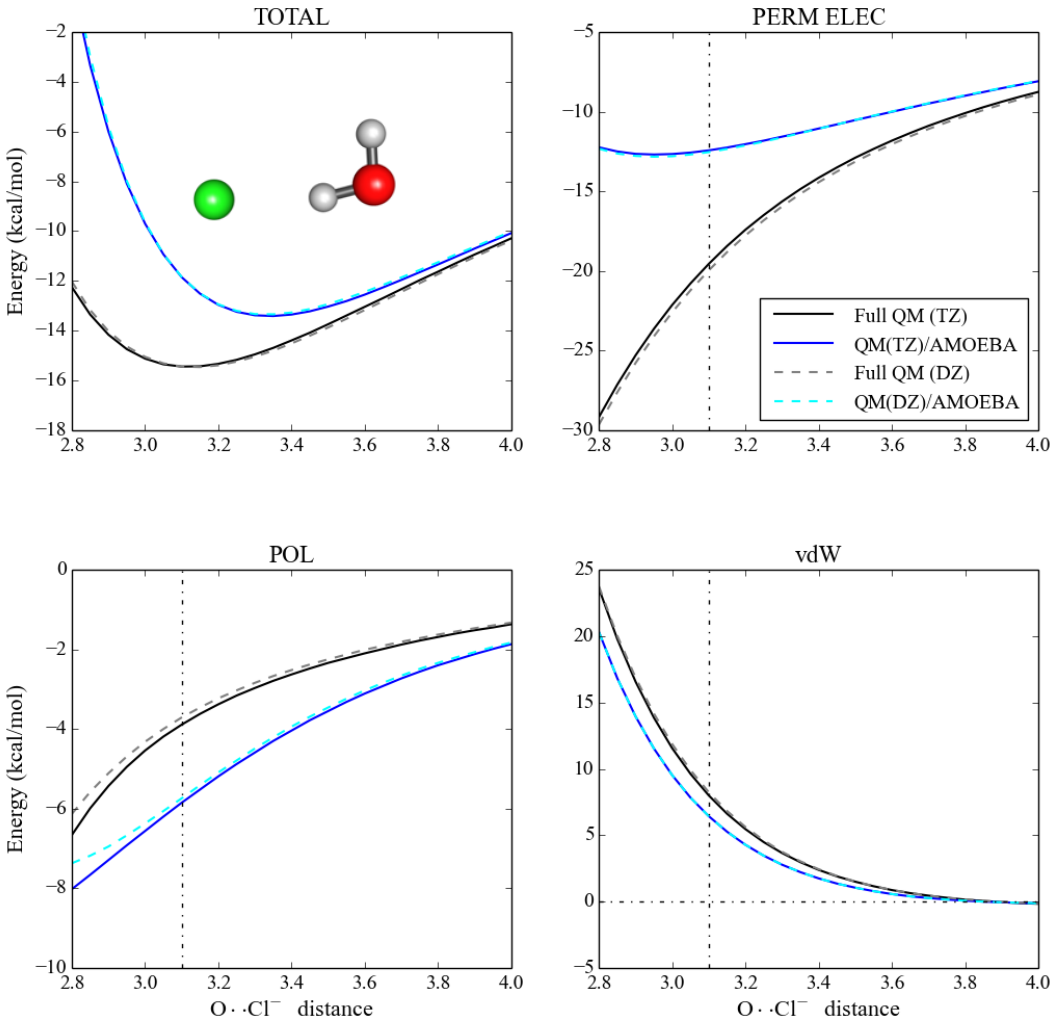

Figure E.4: Total interaction energies and their components (in kcal/mol) for the water-Cl^- complex as functions of the O··Cl^- distance. In the QM/AMOEBA calculations, Cl^- is treated with QM and the H_2O molecule is described by AMOEBA. The other computational and plotting details are the same as in Figure E.3.

Table E.1: Full QM interaction energies and ALMO-EDA results (in kcal/mol) evaluated with def2-SVPD and def2-TZVPPD for three solutes (NH_3, NH_4^+ and CN^-) solvated with increasing numbers (10, 20 and 30) of water molecules. The B97M-V functional is employed for all these calculations, and counterpoise correction is applied to all the resulting interaction energies.

		def2-SVPD				def2-TZVPPD			
		INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
10 H_2O	NH_3	-16.03	-25.37	-6.09	15.43	-16.15	-25.75	-6.07	15.67
	NH_4^+	-51.48	-55.65	-21.43	25.61	-51.97	-55.02	-22.57	25.63
	CN^-	-69.47	-76.29	-17.16	23.98	-69.10	-75.56	-18.25	24.71
20 H_2O	NH_3	-17.80	-26.00	-6.61	14.81	-17.93	-26.36	-6.63	15.06
	NH_4^+	-73.20	-72.30	-26.16	25.26	-73.50	-71.42	-27.37	25.29
	CN^-	-81.99	-84.74	-20.31	23.06	-81.54	-83.88	-21.44	23.77
30 H_2O	NH_3	-18.26	-26.08	-6.65	14.47	-18.35	-26.41	-6.67	14.73
	NH_4^+	-78.48	-76.34	-28.32	26.19	-78.58	-75.16	-29.50	26.07
	CN^-	-90.99	-91.91	-21.77	22.69	-90.20	-90.62	-22.94	23.35

Table E.2: QM/AMOEBA interaction energies and their components (in kcal/mol) for three solutes (NH_3, NH_4^+ and CN^-) solvated with increasing numbers (10, 20 and 30) of water molecules. The B97M-V functional and two basis sets (def2-SVPD and def2-TZVPPD) are employed for the QM region.

		def2-SVPD				def2-TZVPPD			
		INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
10 H_2O	NH_3	-15.88	-15.39	-8.44	7.95	-15.35	-14.90	-8.40	7.95
	NH_4^+	-54.54	-47.79	-18.90	12.15	-54.99	-48.01	-19.13	12.15
	CN^-	-61.08	-53.92	-23.58	16.42	-58.21	-51.13	-23.50	16.42
20 H_2O	NH_3	-17.14	-15.50	-8.94	7.30	-16.59	-14.99	-8.91	7.30
	NH_4^+	-75.38	-64.03	-22.66	11.30	-75.84	-64.24	-22.89	11.30
	CN^-	-73.70	-61.7-	-27.32	15.31	-70.76	-58.78	-27.29	15.31
30 H_2O	NH_3	-17.37	-15.59	-8.92	7.14	-16.84	-15.09	-8.89	7.14
	NH_4^+	-80.02	-66.44	-24.66	11.08	-80.48	-66.65	-24.91	11.08
	CN^-	-81.48	-67.70	-28.75	14.97	-78.54	-64.85	-28.65	14.97

E.3 Original EDA Results for Some Solute-Solvent Systems

Table E.3: Total interaction energies and energy components (in kcal/mol) for the interaction between F^- and 10 surrounding H_2O molecules in 10 isomers. The employed QM level of theory is ωB97X-V/def2-TZVPPD. The same data are plotted in Figure 6.5.

Isomer	Full QM				QM/AMOEBA			
	INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
1	-121.08	-150.75	-41.06	70.73	-98.26	-91.74	-48.18	41.66
2	-121.04	-150.71	-41.05	70.71	-98.26	-91.70	-48.17	41.62
3	-125.54	-154.80	-41.20	70.46	-100.28	-95.05	-49.59	44.36
4	-119.74	-146.78	-40.46	67.50	-98.18	-90.50	-47.74	40.06
5	-119.68	-147.13	-40.48	67.93	-97.67	-90.40	-47.63	40.36
6	-116.97	-146.42	-40.71	70.16	-94.77	-88.59	-47.59	41.41
7	-119.65	-147.28	-40.56	68.18	-97.55	-90.39	-47.70	40.54
8	-121.03	-150.54	-41.09	70.61	-99.38	-92.57	-47.61	40.80
9	-118.26	-147.39	-40.60	69.73	-96.98	-90.14	-47.05	40.21
10	-115.47	-147.98	-41.40	73.91	-92.18	-88.40	-48.39	44.61

Table E.4: Total interaction energies and their components (in kcal/mol) for the interactions between three solutes (NH_3, NH_4^+, CN^-) and increasing numbers (10–100) of solvent water molecules computed with full QM and QM/AMOEBA. The employed level of theory for QM is B97M-V/def2-SVPD. Counterpoise corrections for BSSE are applied to full QM calculations for interaction energies.

NH_3 with	Full QM				QM/AMOEBA			
n H_2O	INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
10	-16.03	-25.37	-6.09	15.43	-15.88	-15.39	-8.44	7.95
20	-17.80	-26.00	-6.61	14.81	-17.14	-15.50	-8.94	7.30
30	-18.26	-26.08	-6.65	14.47	-17.37	-15.59	-8.92	7.14
40	-19.39	-26.75	-6.98	14.34	-18.40	-16.12	-9.32	7.04
50	-17.68	-25.76	-6.48	14.56	-16.29	-14.88	-8.40	6.99
60	-17.40	-25.53	-6.33	14.47	-15.85	-14.63	-8.18	6.95
70	-17.56	-25.61	-6.39	14.44	-15.96	-14.67	-8.22	6.93
80	-16.76	-25.15	-6.23	14.62	-15.02	-14.05	-7.87	6.91
90	-16.56	-25.00	-6.13	14.57	-14.77	-13.92	-7.74	6.90
100	-16.75	-25.13	-6.20	14.58	-14.95	-14.01	-7.82	6.89
NH_4^+ with	Full QM				QM/AMOEBA			
n H_2O	INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
10	-51.48	-55.65	-21.43	25.61	-54.54	-47.79	-18.90	12.15
20	-73.20	-72.30	-26.16	25.26	-75.38	-64.03	-22.66	11.30
30	-78.48	-76.34	-28.32	26.19	-80.02	-66.44	-24.66	11.08
40	-82.57	-79.07	-29.67	26.17	-83.61	-68.75	-25.84	10.99
50	-80.53	-76.12	-30.55	26.14	-80.91	-65.26	-26.59	10.93
60	-86.18	-81.13	-31.20	26.14	-86.76	-70.55	-27.11	10.90
70	-84.65	-79.08	-31.86	26.29	-85.38	-68.53	-27.72	10.87
80	-88.38	-82.40	-32.32	26.33	-89.18	-71.92	-28.12	10.86
90	-92.50	-86.23	-32.70	26.43	-93.38	-75.77	-28.46	10.85
100	-89.04	-82.43	-32.99	26.37	-89.95	-72.06	-28.73	10.84
CN^- with	Full QM				QM/AMOEBA			
n H_2O	INT	ELEC	POL	VDW	INT	ELEC	POL	VDW
10	-69.47	-76.29	-17.16	23.98	-61.08	-53.92	-23.58	16.42
20	-81.99	-84.74	-20.31	23.06	-73.70	-61.70	-27.32	15.31
30	-90.99	-91.91	-21.77	22.69	-81.48	-67.70	-28.75	14.97
40	-102.72	-103.05	-22.35	22.68	-92.88	-78.42	-29.28	14.82
50	-113.35	-112.72	-23.04	22.41	-103.06	-87.80	-30.00	14.74
60	-103.73	-102.48	-23.43	22.18	-93.37	-77.69	-30.35	14.67
70	-105.47	-103.72	-23.84	22.10	-94.77	-78.64	-30.76	14.63
80	-103.85	-101.64	-24.25	22.04	-92.85	-76.29	-31.16	14.60
90	-105.03	-102.52	-24.50	21.99	-94.10	-77.31	-31.37	14.57
100	-106.56	-103.73	-24.78	21.94	-95.63	-78.54	-31.64	14.56

E.4 Preliminary Results of QM/AMOEBA with Gaussian-Blurred Monopoles

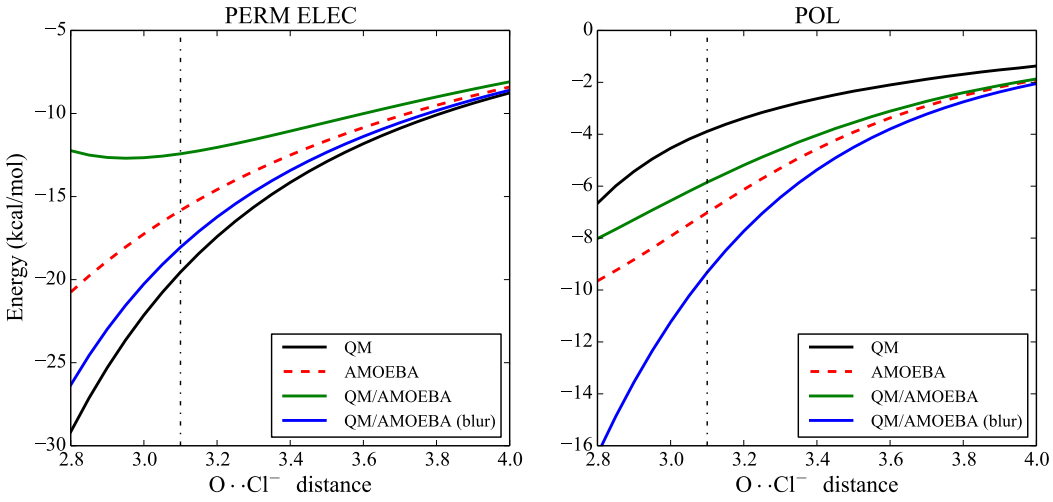

Figure E.5: Permanent electrostatic and polarization energies upon the rigid dissociation of the water-Cl^- complex evaluated with full QM, full AMOEBA, unmodified QM/AMOEBA, and QM/AMOEBA with Gaussian-blurred monopoles. The def2-TZVPPD basis is used for all the full QM and QM/AMOEBA calculations. For the blurred AMOEBA water model, monopoles that correspond to the valence electrons of O and H are replaced by Gaussian functions with exponents for O: 0.66 a_0^{-2}, H: 0.95 a_0^{-2}, while all higher multipole moments and monopoles corresponding to nuclei and core electrons remain unchanged.

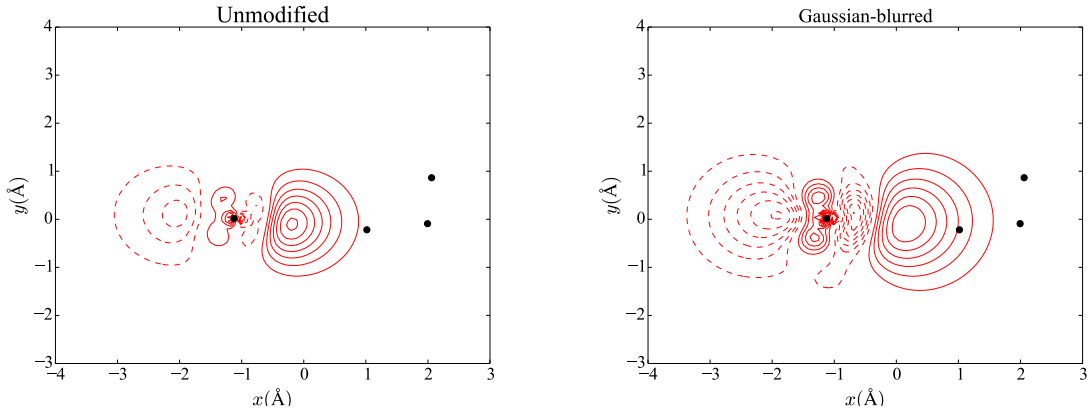

Figure E.6: Illustration of the electron density of Cl^- polarized by unmodified (left) and Gaussian-blurred (right) AMOEBA models. The computational details are the same as in Figure E.5. Values plotted are polarized electron density of Cl^- subtracting that of an isolated, unpolarized Cl^- at the same position, which is integrated to the xy-plane: $\Delta\rho(x,y) = \int \mathrm{d}z\ \Delta\rho(x,y,z)$. The contour lines are evenly spaced at 0.1 $e^-/\text{Å}^3$, where the solid lines indicate density enhancement and dashed lines for density depletion. The black dots indicate the positions of nuclei, where Cl^- is located near the center of each plot. It is manifested that the polarizing effect of AMOEBA water is enhanced once the point monopoles are replaced by Gaussian functions.

E.5 QM/AMOEBA Solute-Solvent Interactions with Growing QM Regions

With the three systems investigated in Sec. 6.3.6 (NH_3, NH_4^+, and CN^- solvated in increasing numbers of water molecules), we compute the solute-solvent interaction energies with QM/AMOEBA for which the QM region includes both the solute and the nearest 10 water molecules. In these calculations, the solute-solvent interaction energies are evaluated as follows:

$$\begin{aligned} E_{\text{int}} = {} & E[\text{QM}(\text{solute} + 10\text{H}_2\text{O})/\text{AMOEBA}] \\ & - E[\text{QM}(10\text{H}_2\text{O})/\text{AMOEBA}] - E[\text{QM}(\text{solute})]. \end{aligned} \tag{E.1}$$

The QM level of theory used here is the same as in Sec. 6.3.6 (B97M-V/def2-SVPD), and the last two terms in Eq. (E.1) are counterpoise-corrected. Here we only look at two different sizes of systems: a solute solvated in 20 H_2O molecules and in 100 H_2O molecules. The resulting solute-solvent interaction energies, as well as the results of the original QM/AMOEBA and full QM calculations (in order to facilitate the comparison), are collected in Table E.5. We see that for NH_3 and CN^- where the original QM(solute only)/MM model underestimates their interactions with the solvent, including the first 10 water molecules into the QM region overall improves the agreement with the full QM solute-solvent interaction energies, while it should be noted that for CN^-, the new QM/AMOEBA results overshoot the full QM answers for both the 20-water and 100-water systems. The situation of NH_4^+, however, is rather different: including the first 10 H_2O molecules into the QM region yields systematically underbound results. For the 100-water system, the resulting solute-solvent interaction energy deviates from the full QM result by almost 6 kcal/mol, which is much poorer compared to the result given by the original QM/AMOEBA calculation ("1/100"). This example indicates that for a solute-solvent system with a given size, simply including several closest solvent molecules into the QM region does not guarantee to give a more accurate solute-solvent interaction energy compared against the full QM result.

In order to find out the underlying reason for the significantly underbound result for the $NH_4^+(H_2O)_{100}$ system, we divide the 100 water molecules into the inner 10 and the outer 90, and then evaluate their interactions with the NH_4^+ solute separately. The results (including the components of these interaction energies) are shown in Table E.6. Looking at these two 2-body interactions first (neglecting the mutual polarization between the inner 10 and the outer 90 water molecules), QM/AMOEBA has well-canceled errors for the interaction of NH_4^+ with the inner 10 (overestimated) and the outer 90 (underestimated) water molecules. For the sum of these two 2-body terms, the net negative error of QM/AMOEBA ("1/100") is in line with its overbound result in Table E.5. Treating the inner 10 water molecules with QM, however, breaks this cancellation of errors and leads to an underestimated solute-solvent interaction (by 2.42 kcal/mol if we consider the 2-body terms only), which partly explains the degraded accuracy. Reintroducing the 3-body effects (e.g., the effect of the mutual polarization between the inner (QM) and outer (AMOEBA) water molecules on their interactions

Table E.5: Interaction energies (in kcal/mol) between three solutes (NH_3, NH_4^+ and CN^-) and the surrounding water molecules (20 or 100) evaluated with QM/AMOEBA (with two distinct partitioning schemes) and full QM. For the QM/AMOEBA calculations, "1/20" denotes that only the solute is treated with QM while all the 20 solvent molecules are described by AMOEBA, and "1+10/10" indicates that the nearest 10 water molecules are also included into the QM region. "1/100" and "1+10/90" are defined in the same manner.

	Solute interacting with 20 H_2O molecules		
	QM/AMOEBA (1/20)	QM/AMOEBA (1+10/10)	Full QM
ammonia	-17.14	-17.91	-17.80
ammonium	-75.38	-71.15	-73.20
cyanide	-73.70	-83.19	-81.99
	Solute interacting with 100 H_2O molecules		
	QM/AMOEBA (1/100)	QM/AMOEBA (1+10/90)	Full QM
ammonia	-14.95	-16.18	-16.75
ammonium	-89.95	-83.14	-89.04
cyanide	-95.63	-108.85	-106.56

Table E.6: Total interaction energies and their components (in kcal/mol) for NH_4^+ interacting with the inner 10 and outer 90 water molecules in the $NH_4^+(H_2O)_{100}$ system.

Interaction between NH_4^+ and the inner 10 water molecules			
	Full QM	QM/AMOEBA	Error vs. QM
PERM ELEC	−55.65	−47.79	7.86
POL	−21.43	−18.90	2.53
VDW	25.61	12.15	−13.46
TOTAL	−51.47	−54.54	−3.07
Interaction between NH_4^+ and the outer 90 water molecules			
	Full QM	QM/AMOEBA	Error vs. QM
PERM ELEC	−27.09	−26.03	1.06
POL	−12.14	−10.86	1.28
VDW	−1.40	−1.31	0.09
TOTAL	−40.63	−38.21	2.42

with the solute) results in a further underestimated interaction energy between NH_4^+ and 100 water molecules, as the overall error of the "1+10/90" QM/AMOEBA calculation is as large as +5.90 kcal/mol. This is most likely due to the unrealistic perturbation to the electron

density of the QM water molecules by the AMOEBA embedding.

From these examples, we see that the consequence of including the solvent molecules that are in close contact with the solute is not as straightforward as one usually expects. It depends on the interplay of long-range and short-range errors for the interaction crossing the QM/MM boundary: if the error for the short-range QM/AMOEBA interaction dominates (e.g. the CN^- case), increasing the size of the QM region is able to reduce the error in total solute-solvent interaction; if the errors for short- and long-range QM/AMOEBA interactions are somewhat balanced (e.g. the NH_4^+ case), then this change in QM/MM partitioning might disrupt the original cancellation of errors and yield a less accurate result. The mutual polarization between the QM and MM solvent molecules can also make a difference to the resulting solute-solvent interaction energy, which can be regarded as a "3-body" effect. While it should be a step towards the right direction, in principle, to refine the treatment of the solvent molecules that are in close contact with the solute, the resulting changes to the QM/MM interface should always be carefully investigated.

References

(1) Slater, J. C. The theory of complex spectra. *Phys. Rev.* **1929**, *34*, 1293.

(2) Condon, E. The theory of complex spectra. *Phys. Rev.* **1930**, *36*, 1121.

(3) Szabo, A.; Ostlund, N. S., *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation: 2012.

(4) Grüneis, A.; Hirata, S.; Ohnishi, Y.-y.; Ten-no, S. Perspective: Explicitly correlated electronic structure theory for complex systems. *J. Chem. Phys.* **2017**, *146*, 080901.

(5) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133.

(6) Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **1951**, *23*, 69.

(7) Löwdin, P.-O. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *J. Chem. Phys.* **1950**, *18*, 365–375.

(8) Pulay, P. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Lett.* **1980**, *73*, 393–398.

(9) Pulay, P. Improved SCF convergence acceleration. *J. Comput. Chem.* **1982**, *3*, 556–560.

(10) Hu, X.; Yang, W. Accelerating self-consistent field convergence with the augmented Roothaan–Hall energy function. *J. Chem. Phys.* **2010**, *132*, 054109.

(11) Kudin, K. N.; Scuseria, G. E.; Cancès, E. A black-box self-consistent field convergence algorithm: One step closer. *J. Chem. Phys.* **2002**, *116*, 8255–8261.

(12) Head-Gordon, M.; Pople, J. A. Optimization of wave function and geometry in the finite basis hartree-fock method. *J. Phys. Chem.* **1988**, *92*, 3063–3069.

(13) Hutter, J.; Parrinello, M.; Vogel, S. Exponential transformation of molecular orbitals. *J. Chem. Phys.* **1994**, *101*, 3862–3865.

(14) Hohenberg, P; Kohn, W Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864.

(15) Wang, Y. A.; Carter, E. A. In *Theoretical methods in condensed phase chemistry*; Springer: 2002, pp 117–184.

(16) Wesolowski, T. A.; Wang, Y. A., *Recent progress in orbital-free density functional theory*; World Scientific: 2013; Vol. 6.

(17) Xia, J.; Carter, E. A. Density-decomposed orbital-free density functional theory for covalently bonded molecules and materials. *Phys. Rev. B* **2012**, *86*, 235109.

(18) Ke, Y.; Libisch, F.; Xia, J.; Wang, L.-W.; Carter, E. A. Angular-momentum-dependent orbital-free density functional theory. *Phys. Rev. Lett.* **2013**, *111*, 066402.

(19) Yao, K.; Parkhill, J. Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks. *J. Chem. Theory Comput.* **2016**, *12*, 1139–1147.

(20) Hung, L.; Carter, E. A. Accurate simulations of metals at the mesoscale: Explicit treatment of 1 million atoms with quantum mechanics. *Chem. Phys. Lett.* **2009**, *475*, 163–170.

(21) Shin, I.; Ramasubramaniam, A.; Huang, C.; Hung, L.; Carter, E. A. Orbital-free density functional theory simulations of dislocations in aluminum. *Philos. Mag.* **2009**, *89*, 3195–3213.

(22) Shin, I.; Carter, E. A. First-principles simulations of plasticity in body-centered-cubic magnesium–lithium alloys. *Acta Mater.* **2014**, *64*, 198–207.

(23) Huang, C.; Carter, E. A. Nonlocal orbital-free kinetic energy density functional for semiconductors. *Phys. Rev. B* **2010**, *81*, 045206.

(24) Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

(25) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(26) Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham density functional theory descending a staircase. *J. Chem. Phys.* **2016**, *145*, 130901.

(27) Klimeš, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* **2012**, *137*, 120901.

(28) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-corrected mean-field electronic structure methods. *Chem. Rev.* **2016**, *116*, 5105–5154.

(29) Hermann, J.; DiStasio Jr, R. A.; Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chem. Rev.* **2017**, *117*, 4714–4758.

(30) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.* **2005**, *123*, 062201.

(31) Dirac, P. A. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 1930; Vol. 26, pp 376–385.

(32) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(33) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244.

(34) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098.

(35) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

(36) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(37) Zhang, Y.; Yang, W. Comment on Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1998**, *80*, 890.

(38) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(39) Becke, A. D. Simulation of delocalized exchange by local density functionals. *J. Chem. Phys* **2000**, *112*, 4020–4026.

(40) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(41) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(42) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(43) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(44) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(45) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. Combining long-range configuration interaction with short-range density functionals. *Chem. Phys. Lett.* **1997**, *275*, 151–160.

(46) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. A long-range correction scheme for generalized-gradient-approximation exchange functionals. *J. Chem. Phys.* **2001**, *115*, 3540–3544.

(47) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(48) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(49) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(50) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618.

(51) Eshuis, H.; Bates, J. E.; Furche, F. Electron correlation methods based on the random phase approximation. *Theor. Chem. Acc.* **2012**, *131*, 1084.

(52) Goerigk, L.; Grimme, S. Double-hybrid density functionals. *WIREs Comput. Mol. Sci.* **2014**, *4*, 576–600.

(53) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(54) Chai, J.-D.; Head-Gordon, M. Long-range corrected double-hybrid density functionals. *J. Chem. Phys.* **2009**, *131*, 174105.

(55) Zhang, Y.; Xu, X.; Goddard, W. A. Doubly hybrid density functional for accurate descriptions of nonbond interactions, thermochemistry, and thermochemical kinetics. *Proc. Natl. Acad. Sci.* **2009**, *106*, 4963–4968.

(56) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(57) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(58) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(59) Schröder, H.; Creon, A.; Schwabe, T. Reformulation of the D3 (Becke–Johnson) Dispersion Correction without Resorting to Higher than C_6 Dispersion Coefficients. *J. Chem. Theory Comput.* **2015**, *11*, 3163–3170.

(60) Smith, D. G.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7*, 2197–2203.

(61) Witte, J.; Mardirossian, N.; Neaton, J. B.; Head-Gordon, M. Assessing DFT-D3 Damping Functions Across Widely Used Density Functionals: Can We Do Better? *J. Chem. Theory Comput.* **2017**, *13*, 2043–2052.

(62) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. *J. Chem. Phys.* **2005**, *122*, 154104.

(63) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *J. Chem. Phys.* **2007**, *127*, 154108.

(64) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(65) Román-Pérez, G.; Soler, J. M. Efficient implementation of a van der Waals density functional: application to double-wall carbon nanotubes. *Phys. Rev. Lett.* **2009**, *103*, 096102.

(66) Sabatini, R.; Gorni, T.; de Gironcoli, S. Nonlocal van der Waals density functional made simple and efficient. *Phys. Rev. B* **2013**, *87*, 041108.

(67) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals density functional for general geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(68) Thonhauser, T.; Cooper, V. R.; Li, S.; Puzder, A.; Hyldgaard, P.; Langreth, D. C. Van der Waals density functional: Self-consistent potential and the nature of the van der Waals bond. *Phys. Rev. B* **2007**, *76*, 125112.

(69) Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-accuracy van der Waals density functional. *Phys. Rev. B* **2010**, *82*, 081101.

(70) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional made simple. *Phys. Rev. Lett.* **2009**, *103*, 063004.

(71) Vydrov, O. A.; van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.* **2010**, *133*, 244103.

(72) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

(73) Mardirossian, N.; Head-Gordon, M. Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V. *J. Chem. Phys.* **2015**, *142*, 074111.

(74) Mardirossian, N.; Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(75) Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **2015**, *11*, 2473–2486.

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

(77) Mardirossian, N.; Head-Gordon, M. Note: The performance of new density functionals for a recent blind test of non-covalent interactions. *J. Chem. Phys.* **2016**, *145*, 186101.

(78) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. The S66×8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905.

(79) Manna, D.; Kesharwani, M. K.; Sylvetsky, N.; Martin, J. M. Conventional and explicitly correlated ab initio benchmark study on water clusters: revision of the BEGDB and WATER27 datasets. *J. Chem. Theory Comput.* **2017**, 3136–3152.

(80) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G.; Vanommeslaeghe, K.; MacKerell Jr, A. D.; Merz Jr, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.* **2017**, *147*, 161727.

(81) Goerigk, L.; Hansen, A.; Bauer, C. A.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo with the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, DOI: DOI:10.1039/c7cp04913g.

(82) James, S. L. Metal-organic frameworks. *Chem. Soc. Rev.* **2003**, *32*, 276–288.

(83) Zhou, H.-C.; Kitagawa, S. Metal–organic frameworks (MOFs). *Chem. Soc. Rev.* **2014**, *43*, 5415–5418.

(84) Lehn, J.-M., *Supramolecular chemistry*; 995; Vch, Weinheim: 1995; Vol. 1.

(85) Steed, J. W.; Atwood, J. L.; Gale, P. A., *Definition and emergence of supramolecular chemistry*; Wiley Online Library: 2012.

(86) Reichardt, C.; Welton, T., *Solvents and solvent effects in organic chemistry*; John Wiley & Sons: 2011.

(87) Witte, J.; Neaton, J. B.; Head-Gordon, M. Push it to the limit: Characterizing the convergence of common sequences of basis sets for intermolecular interactions as described by density functional theory. *J. Chem. Phys.* **2016**, *144*, 194306.

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

(89) Kruse, H.; Goerigk, L.; Grimme, S. Why the standard B3LYP/6-31G* model chemistry should not be used in DFT calculations of molecular thermochemistry: understanding and correcting the problem. *J. Org. Chem.* **2012**, *77*, 10824–10834.

(90) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, 054107.

(91) Witte, J.; Neaton, J. B.; Head-Gordon, M. Effective empirical corrections for basis set superposition error in the def2-SVPD basis: gCP and DFT-C. *J. Chem. Phys.* **2017**, *146*, 234105.

(92) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem. Rev.* **1994**, *94*, 1887–1930.

(93) McDaniel, J. G.; Schmidt, J. Next-Generation Force Fields from Symmetry-Adapted Perturbation Theory. *Annu. Rev. Phys. Chem.* **2016**, *67*, 467–488.

(94) Stone, A., *The Theory of Intermolecular Forces*; Oxford University Press: 2013.

(95) London, F. The general theory of molecular forces. *Trans. Faraday Soc.* **1937**, *33*, 8b–26.

(96) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(97) Horn, P. R.; Mao, Y.; Head-Gordon, M. Probing non-covalent interactions with a second generation energy decomposition analysis using absolutely localized molecular orbitals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 23067–23079.

(98) Phipps, M. J.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chem. Soc. Rev.* **2015**, *44*, 3177–3211.

(99) Pastorczak, E.; Corminboeuf, C. Perspective: Found in translation: Quantum chemical tools for grasping non-covalent interactions. *J. Chem. Phys.* **2017**, *146*, 120901.

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

(101) Stone, A. J.; Misquitta, A. J. Charge-transfer in Symmetry-Adapted Perturbation Theory. *Chem. Phys. Lett.* **2009**, *473*, 201–205.

(102) Misquitta, A. J. Charge Transfer from Regularized Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **2013**, *9*, 5313–5326.

(103) Lao, K. U.; Herbert, J. M. Energy Decomposition Analysis with a Stable Charge-Transfer Term for Interpreting Intermolecular Interactions. *J. Chem. Theory Comput.* **2016**, *12*, 2569–2582.

(104) Kitaura, K.; Morokuma, K. A New Energy Decomposition Scheme for Molecular Interactions within the Hartree-Fock Approximation. *Int. J. Quantum Chem.* **1976**, *10*, 325–340.

(105) Morokuma, K. Why Do Molecules Interact? The Origin of Electron Donor-Acceptor Complexes, Hydrogen Bonding and Proton Affinity. *Acc. Chem. Res.* **1977**, *10*, 294–300.

(106) Ziegler, T.; Rauk, A. On the Calculation of Bonding Energies by the Hartree Fock Slater Method. *Theor. Chem. Acc.* **1977**, *46*, 1–10.

(107) Ziegler, T.; Rauk, A. A Theoretical Study of the Ethylene-Metal Bond in Complexes between Cu^+, Ag^+, Pt^0, or Pt^{2+} and Ethylene, Based on Hartree-Fock-Slater Transition-State Method. *Inorg. Chem.* **1979**, *18*, 1558–1565.

(108) Bickelhaupt, F. M.; Baerends, E. J. In *Reviews in Computational Chemistry*, Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons, Inc.: New York, 2007; Vol. 15; Chapter 1, pp 1–86.

(109) Von Hopffgarten, M.; Frenking, G. Energy decomposition analysis. *WIREs Comput. Mol. Sci.* **2012**, *2*, 43–62.

(110) Mo, Y.; Gao, J.; Peyerimhoff, S. D. Energy decomposition analysis of intermolecular interactions using a block-localized wave function approach. *J. Chem. Phys.* **2000**, *112*, 5530–5538.

(111) Mo, Y.; Song, L.; Lin, Y. Block-localized wavefunction (BLW) method at the density functional theory (DFT) level. *J. Phys. Chem. A* **2007**, *111*, 8291–8301.

(112) Mo, Y.; Bao, P.; Gao, J. Energy decomposition analysis based on a block-localized wavefunction and multistate density functional theory. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6760–6775.

(113) Wu, Q.; Ayers, P. W.; Zhang, Y. Density-based energy decomposition analysis for intermolecular interactions with variationally determined intermediate state energies. *J. Chem. Phys.* **2009**, *131*, 164112.

(114) Wu, Q. Variational nature of the frozen density energy in density-based energy decomposition analysis and its application to torsional potentials. *J. Chem. Phys.* **2014**, *140*, 244109.

(115) Khaliullin, R. Z.; Cobar, E. A.; Lochan, R. C.; Bell, A. T.; Head-Gordon, M. Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals. *J. Phys. Chem. A* **2007**, *111*, 8753–8765.

(116) Horn, P. R.; Sundstrom, E. J.; Baker, T. A.; Head-Gordon, M. Unrestricted absolutely localized molecular orbitals for energy decomposition analysis: Theory and applications to intermolecular interactions involving radicals. *J. Chem. Phys.* **2013**, *138*, 134119.

(117) Horn, P. R.; Head-Gordon, M. Polarization contributions to intermolecular interactions revisited with fragment electric-field response functions. *J. Chem. Phys.* **2015**, *143*, 114111.

(118) Horn, P. R.; Mao, Y.; Head-Gordon, M. Defining the contributions of permanent electrostatics, Pauli repulsion, and dispersion in density functional theory calculations of intermolecular interaction energies. *J. Chem. Phys.* **2016**, *144*, 114107.

(119) Mao, Y.; Horn, P. R.; Head-Gordon, M. Energy decomposition analysis in an adiabatic picture. *Phys. Chem. Chem. Phys.* **2017**, *19*, 5944–5958.

(120) Thirman, J.; Head-Gordon, M. An energy decomposition analysis for second-order Møller–Plesset perturbation theory based on absolutely localized molecular orbitals. *J. Chem. Phys.* **2015**, *143*, 084124.

(121) Thirman, J.; Head-Gordon, M. Efficient Implementation of Energy Decomposition Analysis for Second-Order Møller–Plesset Perturbation Theory and Application to Anion–π Interactions. *J. Phys. Chem. A* **2017**, *121*, 717–728.

(122) Ge, Q.; Mao, Y.; Head-Gordon, M. Energy decomposition analysis for exciplexes using absolutely localized molecular orbitals. *in preparation*.

(123) Frenkel, D.; Smit, B., *Understanding molecular simulation: from algorithms to applications*; Academic press: 2001; Vol. 1.

(124) Payne, M. C.; Teter, M. P.; Allan, D. C.; Arias, T.; Joannopoulos, J. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* **1992**, *64*, 1045.

(125) Feynman, R. P. Forces in molecules. *Phys. Rev.* **1939**, *56*, 340.

(126) Pulay, P. Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules: I. Theory. *Mol. Phys.* **1969**, *17*, 197–204.

(127) Pople, J. A.; Gill, P. M.; Johnson, B. G. Kohn–Sham density-functional theory within a finite basis set. *Chem. Phys. Lett.* **1992**, *199*, 557–560.

(128) Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(129) Lifson, S.; Warshel, A. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *J. Chem. Phys.* **1968**, *49*, 5116–5129.

(130) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–78.

(131) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(132) MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(133) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(134) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(135) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Prot. Chem.* **2003**, *66*, 27–85.

(136) Mackerell, A. D. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(137) Khaliullin, R. Z.; Kühne, T. D. Microscopic properties of liquid water from combined ab initio molecular dynamics and energy decomposition studies. *Phys. Chem. Chem. Phys.* **2013**, *15*, 15746–15766.

(138) Demerdash, O.; Yap, E.-H.; Head-Gordon, T. Advanced potential energy surfaces for condensed phase simulation. *Annu. Rev. Phys. Chem.* **2014**, *65*, 149–174.

(139) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(140) Ren, P.; Wu, C.; Ponder, J. W. Polarizable atomic multipole-based molecular mechanics for organic molecules. *J. Chem. Theory Comput.* **2011**, *7*, 3143–3161.

(141) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(142) Wang, Q.; Rackers, J. A.; He, C.; Qi, R.; Narth, C.; Lagardere, L.; Gresh, N.; Ponder, J. W.; Piquemal, J.-P.; Ren, P. General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *J. Chem. Theory Comput.* **2015**, *11*, 2609–2618.

(143) Rackers, J. A.; Wang, Q.; Liu, C.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. An optimized charge penetration model for use with the AMOEBA force field. *Phys. Chem. Chem. Phys.* **2017**, *19*, 276–291.

(144) Deng, S.; Wang, Q.; Ren, P. Estimating and modeling charge transfer from the SAPT induction energy. *J. Comput. Chem.* **2017**, *38*, 2222–2231.

(145) Liu, C.; Qi, R.; Wang, Q.; Piquemal, J.-P.; Ren, P. Capturing Many-Body Interactions with Classical Dipole Induction Models. *J. Chem. Theory Comput.* **2017**, *13*, 2751–2761.

(146) Qi, R.; Wang, Q.; Ren, P. General van der Waals potential for common organic molecules. *Bioorg. Med. Chem.* **2016**, *24*, 4911–4919.

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

(148) Singh, U. C.; Kollman, P. A. A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl + Cl^-$ exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* **1986**, *7*, 718–730.

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

(150) Gao, J. Hybrid quantum and molecular mechanical simulations: an alternative avenue to solvent effects in organic chemistry. *Acc. Chem. Res.* **1996**, *29*, 298–305.

(151) Lin, H.; Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(152) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(153) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. An effective fragment method for modeling solvent effects in quantum mechanical calculations. *J. Chem. Phys.* **1996**, *105*, 1968–1986.

(154) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. The Effective Fragment Potential Method : A QM-Based MM Approach to Modeling Environmental Effects in Chemistry. *J. Phys. Chem. A* **2001**, *105*, 293–307.

(155) Gordon, M. S.; Smith, Q. A.; Xu, P.; Slipchenko, L. V. Accurate first principles model potentials for intermolecular interactions. *Annu. Rev. Phys. Chem.* **2013**, *64*, 553–578.

(156) Sumowski, C. V.; Ochsenfeld, C. A convergence study of QM/MM isomerization energies with the selected size of the QM region for peptidic systems. *J. Phys. Chem. A* **2009**, *113*, 11734–11741.

(157) Flaig, D.; Beer, M.; Ochsenfeld, C. Convergence of electronic structure with the size of the QM region: example of QM/MM NMR shieldings. *J. Chem. Theory Comput.* **2012**, *8*, 2260–2271.

(158) Isborn, C. M.; Götz, A. W.; Clark, M. A.; Walker, R. C.; Martínez, T. J. Electronic absorption spectra from MM and ab initio QM/MM molecular dynamics: environmental effects on the absorption spectrum of photoactive yellow protein. *J. Chem. Theory Comput.* **2012**, *8*, 5092.

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

(160) Elstner, M.; Porezag, D.; Jungnickel, G; Elsner, J; Haugk, M; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260.

(161) Cui, Q.; Elstner, M. Density functional tight binding: values of semi-empirical methods in an ab initio era. *Phys. Chem. Chem. Phys.* **2014**, *16*, 14368–14377.

(162) Elstner, M.; Seifert, G. Density functional tight binding. *Phil. Trans. R. Soc. A* **2014**, *372*, 20120483.

(163) Babin, V.; Leforestier, C.; Paesani, F. Development of a first principles water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.

(164) Babin, V.; Medders, G. R.; Paesani, F. Development of a first principles water potential with flexible monomers. II: Trimer potential energy surface, third virial coefficient, and small clusters. *J. Chem. Theory Comput.* **2014**, *10*, 1599–1607.

(165) Paesani, F. Getting the right answers for the right reasons: toward predictive molecular simulations of water with many-body potential energy functions. *Acc. Chem. Res.* **2016**, *49*, 1844–1851.

(166) Mao, Y.; Horn, P. R.; Mardirossian, N.; Head-Gordon, T.; Skylaris, C.-K.; Head-Gordon, M. Approaching the basis set limit for DFT calculations using an environment-adapted minimal basis with perturbation theory: Formulation, proof of concept, and a pilot implementation. *J. Chem. Phys.* **2016**, *145*, 044109.

(167) Mao, Y.; Demerdash, O.; Head-Gordon, M.; Head-Gordon, T. Assessing ion-water interactions in the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J. Chem. Theory Comput.* **2016**, *12*, 5422–5437.

(168) Dziedzic, J.; Mao, Y.; Shao, Y.; Ponder, J. W.; Head-Gordon, T.; Head-Gordon, M.; Skylaris, C.-K. TINKTEP: A fully self-consistent, mutually polarizable QM/MM approach based on the AMOEBA force field. *J. Chem. Phys.* **2016**, *145*, 124106.

(169) Mao, Y.; Shao, Y.; Dziedzic, J.; Skylaris, C.-K.; Head-Gordon, T.; Head-Gordon, M. Performance of the AMOEBA water model in the vicinity of QM solutes: A diagnosis using energy decomposition analysis. *J. Chem. Theory Comput.* **2017**, *13*, 1963–1979.

(170) Kohn, W; Becke, A. D.; Parr, R. G. Density functional theory of electronic structure. *J. Phys. Chem.* **1996**, *100*, 12974–12980.

(171) Becke, A. D. Density functional calculations of molecular bond energies. *J. Chem. Phys.* **1986**, *84*, 4524–4529.

(172) Becke, A. D. A new inhomogeneity parameter in density-functional theory. *J. Chem. Phys.* **1998**, *109*, 2092–2098.

(173) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation. *Phys. Rev. Lett.* **1999**, *82*, 2544.

(174) Raymond, K. S.; Wheeler, R. A. Compatibility of correlation-consistent basis sets with a hybrid Hartree–Fock/density functional method. *J. Comput. Chem.* **1999**, *20*, 207–216.

(175) Boese, A. D.; Martin, J. M.; Handy, N. C. The role of the basis set: Assessing density functional theory. *J. Chem. Phys.* **2003**, *119*, 3005–3014.

(176) Wang, N. X.; Wilson, A. K. The behavior of density functionals with respect to basis set. I. The correlation consistent basis sets. *J. Chem. Phys.* **2004**, *121*, 7632–7646.

(177) Wang, N. X.; Wilson, A. K. Behaviour of density functionals with respect to basis set: II. Polarization consistent basis sets. *Mol. Phys.* **2005**, *103*, 345–358.

(178) Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(179) Woon, D. E.; Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

(180) Řezáč, J.; Hobza, P. Describing noncovalent interactions beyond the common approximations: How accurate is the gold standard, CCSD(T) at the complete basis set limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155.

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

(182) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.

(183) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010**, *133*, 134105.

(184) Van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G.; van Lenthe, J. H. State of the art in counterpoise theory. *Chem. Rev.* **1994**, *94*, 1873–1885.

(185) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(186) Haunschild, R.; Klopper, W. New accurate reference energies for the G2/97 test set. *J. Chem. Phys.* **2012**, *136*, 164102.

(187) Adamson, R. D.; Gill, P. M.; Pople, J. A. Empirical density functionals. *Chem. Phys. Lett.* **1998**, *284*, 6–11.

(188) White, C. A.; Johnson, B. G.; Gill, P. M.; Head-Gordon, M. The continuous fast multipole method. *Chem. Phys. Lett.* **1994**, *230*, 8–16.

(189) White, C. A.; Johnson, B. G.; Gill, P. M.; Head-Gordon, M. Linear scaling density functional calculations via the continuous fast multipole method. *Chem. Phys. Lett.* **1996**, *253*, 268–278.

(190) White, C. A.; Head-Gordon, M. A J matrix engine for density functional theory calculations. *J. Chem. Phys.* **1996**, *104*, 2620–2629.

(191) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Achieving linear scaling for the electronic quantum Coulomb problem. *Science* **1996**, *271*, 51.

(192) Challacombe, M.; Schwegler, E. Linear scaling computation of the Fock matrix. *J. Chem. Phys.* **1997**, *106*, 5526–5536.

(193) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. Linear and sublinear scaling formation of Hartree–Fock-type exchange matrices. *J. Chem. Phys.* **1998**, *109*, 1663–1669.

(194) Lippert, G.; Hutter, J.; Parrinello, M. The Gaussian and augmented-plane-wave density functional method for ab initio molecular dynamics simulations. *Theor. Chem. Acc.* **1999**, *103*, 124–140.

(195) Dziedzic, J; Hill, Q; Skylaris, C.-K. Linear-scaling calculation of Hartree-Fock exchange energy with non-orthogonal generalised Wannier functions. *J. Chem. Phys.* **2013**, *139*, 214103.

(196) Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **1999**, *71*, 1085.

(197) Bowler, D.; Miyazaki, T O(N) methods in electronic structure calculations. *Rep. Prog. Phys.* **2012**, *75*, 036503.

(198) VandeVondele, J.; Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **2007**, *127*, 114105.

(199) Maslen, P. E.; Ochsenfeld, C.; White, C. A.; Lee, M. S.; Head-Gordon, M. Locality and sparsity of ab initio one-particle density matrices and localized orbitals. *J. Phys. Chem. A* **1998**, *102*, 2215–2222.

(200) Berghold, G.; Parrinello, M.; Hutter, J. Polarized atomic orbitals for linear scaling methods. *J. Chem. Phys.* **2002**, *116*, 1800–1810.

(201) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary basis sets to approximate Coulomb potentials. *Chem. Phys. Lett.* **1995**, *240*, 283–290.

(202) Früchtl, H. A.; Kendall, R. A.; Harrison, R. J.; Dyall, K. G. An implementation of RI–SCF on parallel computers. *Int. J. Quantum Chem.* **1997**, *64*, 63–69.

(203) Weigend, F. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

(204) Sodt, A.; Subotnik, J. E.; Head-Gordon, M. Linear scaling density fitting. *J. Chem. Phys.* **2006**, *125*, 194109.

(205) Sodt, A.; Head-Gordon, M. Hartree-Fock exchange computed using the atomic resolution of the identity approximation. *J. Chem. Phys.* **2008**, *128*, 104106.

(206) Sierka, M.; Hogekamp, A.; Ahlrichs, R. Fast evaluation of the Coulomb potential for electron densities using multipole accelerated resolution of identity approximation. *J. Chem. Phys.* **2003**, *118*, 9136–9148.

(207) Manzer, S. F.; Epifanovsky, E.; Head-Gordon, M. Efficient implementation of the pair atomic resolution of the identity approximation for exact exchange for hybrid and range-separated density functionals. *J. Chem. Theory Comput.* **2015**, *11*, 518–527.

(208) Manzer, S.; Horn, P. R.; Mardirossian, N.; Head-Gordon, M. Fast, accurate evaluation of exact exchange: The occ-RI-K algorithm. *J. Chem. Phys.* **2015**, *143*, 024113.

(209) Jurgens-Lutovsky, R.; Almlöf, J. Dual basis sets in calculations of electron correlation. *Chem. Phys. Lett.* **1991**, *178*, 451–454.

(210) Wolinski, K.; Pulay, P. Second-order Møller–Plesset calculations with dual basis sets. *J. Chem. Phys.* **2003**, *118*, 9497–9503.

(211) Liang, W.; Head-Gordon, M. Approaching the basis set limit in density functional theory calculations using dual basis sets without diagonalization. *J. Phys. Chem. A* **2004**, *108*, 3206–3210.

(212) Steele, R. P.; DiStasio Jr, R. A.; Shao, Y.; Kong, J.; Head-Gordon, M. Dual-basis second-order Møller-Plesset perturbation theory: A reduced-cost reference for correlation calculations. *J. Chem. Phys.* **2006**, *125*, 074108.

(213) Steele, R. P.; Shao, Y.; DiStasio, R. A.; Head-Gordon, M. Dual-basis analytic gradients. 1. Self-consistent field theory. *J. Phys. Chem. A* **2006**, *110*, 13915–13922.

(214) Steele, R. P.; Head-Gordon, M. Dual-basis self-consistent field methods: 6-31G* calculations with a minimal 6-4G primary basis. *Mol. Phys.* **2007**, *105*, 2455–2473.

(215) Steele, R. P.; DiStasio Jr, R. A.; Head-Gordon, M. Non-covalent interactions with dual-basis methods: Pairings for augmented basis sets. *J. Chem. Theory Comput.* **2009**, *5*, 1560–1572.

(216) Deng, J.; Gilbert, A. T.; Gill, P. M. Approaching the Hartree–Fock limit by perturbative methods. *J. Chem. Phys.* **2009**, *130*, 231101.

(217) Deng, J.; Gilbert, A. T.; Gill, P. M. Hartree–Fock perturbative corrections for total and reaction energies. *J. Chem. Phys.* **2010**, *133*, 044116.

(218) Deng, J.; Gilbert, A. T.; Gill, P. M. Density functional triple jumping. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10759–10765.

(219) Mulliken, R. Criteria for the Construction of Good Self-Consistent-Field Molecular Orbital Wave Functions, and the Significance of LCAO-MO Population Analysis. *J. Chem. Phys.* **1962**, *36*, 3428–3439.

(220) Davidson, E. R. Electronic population analysis of molecular wavefunctions. *J. Chem. Phys.* **1967**, *46*, 3320–3324.

(221) Roby, K. R. Quantum theory of chemical valence concepts: I. Definition of the charge on an atom in a molecule and of occupation numbers for electron density shared between atoms. *Mol. Phys.* **1974**, *27*, 81–104.

(222) Heinzmann, R.; Ahlrichs, R. Population analysis based on occupation numbers of modified atomic orbitals (MAOs). *Theor. Chim. Acta.* **1976**, *42*, 33–45.

(223) Ehrhardt, C.; Ahlrichs, R. Population analysis based on occupation numbers II. Relationship between shared electron numbers and bond energies and characterization of hypervalent contributions. *Theor. Chim. Acta.* **1985**, *68*, 231–245.

(224) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.* **1988**, *88*, 899–926.

(225) Lee, M. S.; Head-Gordon, M. Extracting polarized atomic orbitals from molecular orbital calculations. *Int. J. Quantum Chem.* **2000**, *76*, 169–184.

(226) Mayer, I Orthogonal effective atomic orbitais in the topological theory of atoms. *Can. J. Chem.* **1996**, *74*, 939–942.

(227) Cioslowski, J.; Liashenko, A. Atomic orbitals in molecules. *J. Chem. Phys.* **1998**, *108*, 4405–4412.

(228) Lu, W.; Wang, C.; Schmidt, M.; Bytautas, L; Ho, K.; Ruedenberg, K Molecule intrinsic minimal basis sets. I. Exact resolution of ab initio optimized molecular orbitals in terms of deformed atomic minimal-basis orbitals. *J. Chem. Phys.* **2004**, *120*, 2629–2637.

(229) Laikov, D. N. Intrinsic minimal atomic basis representation of molecular electronic wavefunctions. *Int. J. Quantum Chem.* **2011**, *111*, 2851–2867.

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

(231) Adams, W. H. On the Solution of the Hartree-Fock Equation in Terms of Localized Orbitals. *J. Chem. Phys.* **1961**, *34*, 89–102.

(232) Adams, W. H. Orbital theories of electronic structure. *J. Chem. Phys.* **1962**, *37*, 2009–2018.

(233) Stoll, H.; Wagenblast, G.; Preuß, H. On the Use of Local Basis Sets for Localized Molecular Orbitals. *Theor. Chem. Acc.* **1980**, *57*, 169–178.

(234) Khaliullin, R. Z.; Head-Gordon, M.; Bell, A. T. An efficient self-consistent field method for large systems of weakly interacting components. *J. Chem. Phys.* **2006**, *124*, 204105.

(235) Adams, W. Distortion of interacting atoms and ions. *Chem. Phys. Lett.* **1971**, *12*, 295–298.

(236) Lee, M. S.; Head-Gordon, M. Polarized atomic orbitals for self-consistent field electronic structure calculations. *J. Chem. Phys.* **1997**, *107*, 9085–9095.

(237) Lee, M. S.; Head-Gordon, M. Absolute and relative energies from polarized atomic orbital self-consistent field calculations and a second order correction.: Convergence with size and composition of the secondary basis. *Computers and chemistry* **2000**, *24*, 295–301.

(238) Ozaki, T. Variationally optimized atomic orbitals for large-scale electronic structures. *Phys. Rev. B* **2003**, *67*, 155108.

(239) Ozaki, T; Kino, H Numerical atomic basis orbitals from H to Kr. *Phys. Rev. B* **2004**, *69*, 195113.

(240) Junquera, J.; Paz, Ó.; Sánchez-Portal, D.; Artacho, E. Numerical atomic orbitals for linear-scaling calculations. *Phys. Rev. B* **2001**, *64*, 235111.

(241) Basanta, M.; Dappe, Y.; Jelínek, P; Ortega, J Optimized atomic-like orbitals for first-principles tight-binding molecular dynamics. *Comput. Mater. Sci.* **2007**, *39*, 759–766.

(242) Bowler, D.; Miyazaki, T Calculations for millions of atoms with density functional theory: linear scaling shows its potential. *J. Phys.: Condens. Matter* **2010**, *22*, 074207.

(243) Bowler, D.; Miyazaki, T; Gillan, M. Recent progress in linear scaling ab initio electronic structure techniques. *J. Phys.: Condens. Matter* **2002**, *14*, 2781.

(244) Torralba, A.; Todorović, M; Brázdová, V; Choudhury, R; Miyazaki, T; Gillan, M.; Bowler, D. Pseudo-atomic orbitals as basis sets for the O(N) DFT code CONQUEST. *J. Phys.: Condens. Matter* **2008**, *20*, 294206.

(245) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.* **2005**, *122*, 084119.

(246) Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. Recent progress in linear-scaling density functional calculations with plane waves and pseudopotentials: the ONETEP code. *J. Phys.: Condens. Matter* **2008**, *20*, 064209.

(247) Skylaris, C.-K.; Mostofi, A. A.; Haynes, P. D.; Diéguez, O.; Payne, M. C. Nonorthogonal generalized Wannier function pseudopotential plane-wave method. *Phys. Rev. B* **2002**, *66*, 035119.

(248) Mostofi, A. A.; Haynes, P. D.; Skylaris, C.-K.; Payne, M. C. Preconditioned iterative minimization for linear-scaling electronic structure calculations. *J. Chem. Phys.* **2003**, *119*, 8842–8848.

(249) Rayson, M.; Briddon, P. Highly efficient method for Kohn-Sham density functional calculations of 500–10000 atom systems. *Phys. Rev. B* **2009**, *80*, 205104.

(250) Rayson, M. Rapid filtration algorithm to construct a minimal basis on the fly from a primitive Gaussian basis. *Comput. Phys. Commun.* **2010**, *181*, 1051–1056.

(251) Nakata, A.; Bowler, D. R.; Miyazaki, T. Efficient Calculations with Multisite Local Orbitals in a large-scale DFT Code CONQUEST. *J. Chem. Theory Comput.* **2014**, *10*, 4813–4822.

(252) Lin, L.; Lu, J.; Ying, L.; Weinan, E Adaptive local basis set for Kohn–Sham density functional theory in a discontinuous Galerkin framework I: Total energy calculation. *J. Comput. Phys.* **2012**, *231*, 2140–2154.

(253) Lin, L.; Lu, J.; Ying, L.; Weinan, E Optimized local basis set for Kohn–Sham density functional theory. *J. Comput. Phys.* **2012**, *231*, 4515–4529.

(254) Head-Gordon, M.; Maslen, P. E.; White, C. A. A tensor formulation of many-electron theory in a nonorthogonal single-particle basis. *J. Chem. Phys.* **1998**, *108*, 616–625.

(255) Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184–215.

(256) Nocedal, J.; Wright, S., *Numerical optimization*; Springer Science & Business Media: 2006.

(257) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528.

(258) Gill, P. M.; Johnson, B. G.; Pople, J. A. A standard grid for density functional calculations. *Chem. Phys. Lett.* **1993**, *209*, 506–512.

(259) Van Voorhis, T.; Head-Gordon, M. A geometric approach to direct minimization. *Mol. Phys.* **2002**, *100*, 1713–1721.

(260) Sun, J.; Haunschild, R.; Xiao, B.; Bulik, I. W.; Scuseria, G. E.; Perdew, J. P. Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence. *J. Chem. Phys.* **2013**, *138*, 044113.

(261) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(262) Jensen, F. Polarization consistent basis sets: principles. *J. Chem. Phys.* **2001**, *115*, 9113–9125.

(263) Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn–Sham basis set limit. *J. Chem. Phys.* **2002**, *116*, 7372–7379.

(264) Jensen, F. Polarization consistent basis sets. III. The importance of diffuse functions. *J. Chem. Phys.* **2002**, *117*, 9234–9240.

(265) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.

(266) Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *J. Chem. Phys.* **1984**, *80*, 3265–3269.

(267) Karton, A.; Daon, S.; Martin, J. M. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **2011**, *510*, 165–178.

(268) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first-and second-row compounds. *J. Chem. Phys.* **1991**, *94*, 7221–7230.

(269) Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals-Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

(270) Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark database of barrier heights for heavy atom transfer, nucleophilic substitution, association, and unimolecular reactions and its use to test theoretical methods. *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

(271) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(272) Martin, J. M. What can we learn about dispersion from the conformer surface of *n*-pentane? *J. Phys. Chem. A* **2013**, *117*, 3118–3132.

(273) Jurečka, P.; Šponer, J.; Černỳ, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(274) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, $\delta_{\mathrm{MP2}}^{\mathrm{CCSD(T)}}$: Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. *J. Chem. Phys.* **2011**, *135*, 194102.

(275) Řezáč, J.; Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

(276) Lao, K. U.; Herbert, J. M. An improved treatment of empirical dispersion and a many-body energy decomposition scheme for the explicit polarization plus symmetry-adapted perturbation theory (XSAPT) method. *J. Chem. Phys.* **2013**, *139*, 034107.

(277) Sedlak, R.; Janowski, T.; Pitonak, M.; Rezac, J.; Pulay, P.; Hobza, P. Accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

(278) Strout, D. L.; Scuseria, G. E. A quantitative study of the scaling properties of the Hartree–Fock method. *J. Chem. Phys.* **1995**, *102*, 8448–8452.

(279) Aubke, F; Wang, C Carbon monoxide as a σ-donor ligand in coordination chemistry. *Coord. Chem. Rev.* **1994**, *137*, 483–524.

(280) Lupinetti, A. J.; Frenking, G.; Strauss, S. H. Nonclassical metal carbonyls: Appropriate definitions with a theoretical justification. *Angew. Chem., Int. Ed.* **1998**, *37*, 2113–2116.

(281) Lupinetti, A. J.; Strauss, S. H.; Frenking, G. Nonclassical metal carbonyls. *Prog. Inorg. Chem.* **2001**, *49*, 1–112.

(282) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. Electron donation in the water–water hydrogen bond. *Chem.-Eur. J.* **2009**, *15*, 851–855.

(283) De Silva, P.; Korchowiec, J. Energy partitioning scheme based on self-consistent method for subsystems: Populational space approach. *J. Comput. Chem.* **2011**, *32*, 1054–1064.

(284) Azar, R. J.; Head-Gordon, M. An energy decomposition analysis for intermolecular interactions from an absolutely localized molecular orbital reference at the coupled-cluster singles and doubles level. *J. Chem. Phys.* **2012**, *136*, 024103.

(285) Tafipolsky, M. Challenging Dogmas: Hydrogen Bond Revisited. *J. Phys. Chem. A* **2016**, *120*, 4550–4559.

(286) Bagus, P. S.; Hermann, K; Bauschlicher Jr, C. W. A new analysis of charge transfer and polarization for ligand–metal bonding: Model studies of Al_4CO and Al_4NH_3. *J. Chem. Phys.* **1984**, *80*, 4378–4386.

(287) Stevens, W. J.; Fink, W. H. Frozen fragment reduced variational space analysis of hydrogen bonding interactions. Application to the water dimer. *Chem. Phys. Lett.* **1987**, *139*, 15–22.

(288) Mitoraj, M. P.; Michalak, A.; Ziegler, T. A Combined Charge and Energy Decomposition Scheme for Bond Analysis. *J. Chem. Theory Comput.* **2009**, *5*, 962–975.

(289) Glendening, E. D.; Streitwieser, A. Natural energy decomposition analysis: An energy partitioning procedure for molecular Interactions with application to weak hydrogen bonding, strong ionic, and moderate donor-acceptor interactions. *J. Chem. Phys.* **1994**, *100*, 2900–2909.

(290) Glendening, E. D. Natural energy decomposition analysis: Extension to density functional methods and analysis of cooperative effects in water clusters. *J. Phys. Chem. A* **2005**, *109*, 11936–11940.

(291) Su, P.; Li, H. Energy decomposition analysis of covalent bonds and intermolecular interactions. *J. Chem. Phys.* **2009**, *131*, 014102.

(292) Su, P.; Jiang, Z.; Chen, Z.; Wu, W. Energy Decomposition Scheme Based on the Generalized Kohn–Sham Scheme. *J. Chem. Phys. A* **2014**, *118*, 2531–2542.

(293) Wu, Q.; Van Voorhis, T. Direct optimization method to study constrained systems within density-functional theory. *Phys. Rev. A*, *72*, 024502.

(294) Kaduk, B.; Kowalczyk, T.; Van Voorhis, T. Constrained density functional theory. *Chem. Rev.* **2011**, *112*, 321–370.

(295) Yamabe, S.; Morokuma, K. Molecular orbital studies of hydrogen bonds. IX. Electron distribution analysis. *J. Am. Chem. Soc.* **1975**, *97*, 4458–4465.

(296) Weinhold, F.; Klein, R. A. What is a hydrogen bond? Mutually consistent theoretical and experimental criteria for characterizing H-bonding interactions. *Mol. Phys.* **2012**, *110*, 565–579.

(297) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.* **1988**, *88*, 899–926.

(298) Mo, Y. Geometrical optimization for strictly localized structures. *J. Chem. Phys.* **2003**, *119*, 1300–1306.

(299) Wang, C.; Guan, L.; Danovich, D.; Shaik, S.; Mo, Y. The origins of the directionality of noncovalent intermolecular interactions. *J. Comput. Chem.* **2016**, *37*, 34–45.

(300) Ramos-Cordoba, E.; Lambrecht, D. S.; Head-Gordon, M. Charge-transfer and the hydrogen bond: Spectroscopic and structural implications from electronic structure calculations. *Faraday Discuss.* **2011**, *150*, 345–362.

(301) Gianinetti, E.; Raimondi, M.; Tornaghi, E. Modification of the Roothaan Equations to Exclude BSSE from Molecular Interaction Calculations. *Int. J. Quantum Chem.* **1996**, *60*, 157–166.

(302) Handy, N. C.; Schaefer III, H. F. On the evaluation of analytic energy derivatives for correlated wave functions. *J. Chem. Phys.* **1984**, *81*, 5031–5033.

(303) Famulari, A.; Gianinetti, E.; Raimondi, M.; Sironi, M. Implementation of gradient-optimization algorithms and force constant computations in BSSE-free direct and conventional SCF approaches. *Int. J. Quantum Chem.* **1998**, *69*, 151–158.

(304) Sharada, S. M.; Bell, A. T.; Head-Gordon, M. A finite difference Davidson procedure to sidestep full ab initio hessian calculation: Application to characterization of stationary points and transition state searches. *J. Chem. Phys.* **2014**, *140*, 164115.

(305) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. Analysis of charge transfer effects in molecular complexes based on absolutely localized molecular orbitals. *J. Chem. Phys.* **2008**, *128*, 184112.

(306) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer III, H. F.; Quack, M. Anchoring the water dimer potential energy surface with explicitly correlated computations and focal point analyses. *J. Chem. Phys.* **2002**, *116*, 690–701.

(307) Ronca, E.; Belpassi, L.; Tarantelli, F. A quantitative view of charge transfer in the hydrogen bond: the water dimer case. *ChemPhysChem* **2014**, *15*, 2682–2687.

(308) Lu, Z.; Zhou, N.; Wu, Q.; Zhang, Y. Directional dependence of hydrogen bonds: A density-based energy decomposition analysis and its implications on force field development. *J. Chem. Theory Comput.* **2011**, *7*, 4038–4049.

(309) Jiao, D.; King, C.; Grossfield, A.; Darden, T. A.; Ren, P. Simulation of Ca^{2+} and Mg^{2+} solvation using polarizable atomic multipole potential. *J. Phys. Chem. B* **2006**, *110*, 18553–18559.

(310) Piquemal, J.-P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. Towards accurate solvation dynamics of divalent cations in water using the polarizable amoeba force field: From energetics to structure. *J. Chem. Phys.* **2006**, *125*, 054511.

(311) Rayón, V. M.; Frenking, G. Bis (benzene) chromium is a δ-bonded molecule and ferrocene is a π-bonded molecule. *Organometallics* **2003**, *22*, 3304–3308.

(312) Swart, M. Metal–ligand bonding in metallocenes: Differentiation between spin state, electrostatic and covalent bonding. *Inorg. Chim. Acta* **2007**, *360*, 179–189.

(313) Stein, C. J.; von Burg, V.; Reiher, M. The Delicate Balance of Static and Dynamic Electron Correlation. *J. Chem. Theory Comput.* **2016**, *12*, 3764–3773.

(314) Coriani, S.; Haaland, A.; Helgaker, T.; Jørgensen, P. The equilibrium structure of ferrocene. *ChemPhysChem* **2006**, *7*, 245–249.

(315) Swart, M.; Ehlers, A. W.; Lammertsma, K. Performance of the OPBE exchange-correlation functional. *Mol. Phys.* **2004**, *102*, 2467–2474.

(316) Azar, R. J.; Horn, P. R.; Sundstrom, E. J.; Head-Gordon, M. Useful lower limits to polarization contributions to intermolecular interactions using a minimal basis of localized orthogonal orbitals: Theory and analysis of the water dimer. *J. Chem. Phys.* **2013**, *138*, 084102.

(317) Marcus, R. A. Electron transfer reactions in chemistry. Theory and experiment. *Rev. Mod. Phys.* **1993**, *65*, 599.

(318) Newton, M. D. Quantum chemical probes of electron-transfer kinetics: the nature of donor-acceptor interactions. *Chem. Rev.* **1991**, *91*, 767–792.

(319) Hammerum, S. Alkyl radicals as hydrogen bond acceptors: Computational evidence. *J. Am. Chem. Soc.* **2009**, *131*, 8627–8635.

(320) Mizuse, K.; Hasegawa, H.; Mikami, N.; Fujii, A. Infrared and Electronic Spectroscopy of Benzene-Ammonia Cluster Radical Cations $[C_6H_6(NH_3)_{1,2}]^+$: Observation of Isolated and Microsolvated σ-Complexes. *J. Phys. Chem. A* **2010**, *114*, 11060–11069.

(321) Attah, I. K.; Platt, S. P.; Meot-Ner, M.; El-Shall, M. S.; Peverati, R.; Head-Gordon, M. What is the structure of the naphthalene–benzene heterodimer radical cation? Binding energy, charge delocalization, and unexpected charge-transfer interaction in stacked dimer and trimer radical cations. *J. Phys. Chem. Lett.* **2015**, *6*, 1111–1118.

(322) Peverati, R.; Platt, S. P.; Attah, I. K.; Aziz, S. G.; El-Shall, M. S.; Head-Gordon, M. Nucleophilic Aromatic Addition in Ionizing Environments: Observation and Analysis of New C–N Valence Bonds in Complexes between Naphthalene Radical Cation and Pyridine. *J. Am. Chem. Soc.* **2017**, *139*, 11923–11932.

(323) Buckingham, A.; Del Bene, J.; McDowell, S. The hydrogen bond. *Chem. Phys. Lett.* **2008**, *463*, 1–10.

(324) Roncaratti, L.; Belpassi, L; Cappelletti, D; Pirani, F; Tarantelli, F Molecular-beam scattering experiments and theoretical calculations probing charge transfer in weakly bound complexes of water. *J. Phys. Chem. A* **2009**, *113*, 15223–15232.

(325) Belpassi, L.; Reca, M. L.; Tarantelli, F.; Roncaratti, L. F.; Pirani, F.; Cappelletti, D.; Faure, A.; Scribano, Y. Charge-Transfer Energy in the Water- Hydrogen Molecular Aggregate Revealed by Molecular-Beam Scattering Experiments, Charge Displacement Analysis, and ab Initio Calculations. *J. Am. Chem. Soc.* **2010**, *132*, 13046–13058.

(326) Cappelletti, D.; Ronca, E.; Belpassi, L.; Tarantelli, F.; Pirani, F. Revealing charge-transfer effects in gas-phase water chemistry. *Acc. Chem. Res.* **2012**, *45*, 1571–1580.

(327) Li, H.; Gordon, M. S.; Jensen, J. H. Charge transfer interaction in the effective fragment potential method. *J. Chem. Phys.* **2006**, *124*, 214108.

(328) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. Anisotropic, polarizable molecular mechanics studies of inter- and intramolecular interactions and ligand-macromolecule complexes. A bottom-up strategy. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.

(329) Lee, A. J.; Rick, S. W. The effects of charge transfer on the properties of liquid water. *J. Chem. Phys.* **2011**, *134*, 184507.

(330) Khoury, L.; Naseem-Khan, S.; Kwapien, K.; Hobaika, Z.; Maroun, R.; Piquemal, J.-P.; Gresh, N. Importance of explicit smeared lone-pairs in anisotropic polarizable molecular mechanics. Torture track angular tests for exchange-repulsion and charge transfer contributions. *J. Comput. Chem.* **2017**, *121*, 3997–1920.

(331) Reed, A. E.; Weinhold, F. Natural bond orbital analysis of near-Hartree–Fock water dimer. *J. Chem. Phys.* **1983**, *78*, 4066–4073.

(332) Glendening, E. D. Natural energy decomposition analysis: Explicit evaluation of electrostatic and polarization effects with application to aqueous clusters of alkali metal cations and neutrals. *J. Am. Chem. Soc.* **1996**, *118*, 2473–2482.

(333) Stone, A. J. Natural Bond Orbitals and the Nature of the Hydrogen Bond. *J. Phys. Chem. A* **2017**, *121*, 1531–1534.

(334) Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *WIREs Comput. Mol. Sci.* **2012**, *2*, 254–272.

(335) Stone, A. J. Computation of charge-transfer energies by perturbation theory. *Chem. Phys. Lett.* **1993**, *211*, 101–109.

(336) Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.* **2005**, *122*, 214109.

(337) Cobar, E. A.; Horn, P. R.; Bergman, R. G.; Head-Gordon, M. Examination of the hydrogen-bonding networks in small water clusters (n= 2–5, 13, 17) using absolutely localized molecular orbital energy decomposition analysis. *Phys. Chem. Chem. Phys.* **2012**, *14*, 15328–15339.

(338) Bell, F.; Ruan, Q. N.; Golan, A.; Horn, P. R.; Ahmed, M.; Leone, S. R.; Head-Gordon, M. Dissociative photoionization of glycerol and its dimer occurs predominantly via a ternary hydrogen-bridged ion–molecule complex. *J. Am. Chem. Soc.* **2013**, *135*, 14229–14239.

(339) Wu, Q.; Van Voorhis, T. Constrained density functional theory and its application in long-range electron transfer. *J. Chem. Theory Comput.* **2006**, *2*, 765–774.

(340) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chem. Acc.* **1977**, *44*, 129–138.

(341) Řezáč, J.; de la Lande, A. Robust, basis-set independent method for the evaluation of charge-transfer energy in noncovalent complexes. *J. Chem. Theory Comput.* **2015**, *11*, 528–537.

(342) Řezáč, J.; de la Lande, A. On the role of charge transfer in halogen bonding. *Phys. Chem. Chem. Phys.* **2017**, *19*, 791–803.

(343) Reed, A. E.; Weinhold, F.; Curtiss, L. A.; Pochatko, D. J. Natural bond orbital analysis of molecular interactions: Theoretical studies of binary complexes of HF, H_2O, NH_3, N_2, O_2, F_2, CO, and CO_2 with HF, H_2O, and NH_3. *J. Chem. Phys.* **1986**, *84*, 5687–5705.

(344) Sokalski, W.; Hariharan, P.; Kaufman, J. J. A self-consistent field interaction energy decomposition study of 12 hydrogen-bonded dimers. *J. Phys. Chem.* **1983**, *87*, 2803–2810.

(345) Hayes, I.; Stone, A. An intermolecular perturbation theory for the region of moderate overlap. *Mol. Phys.* **1984**, *53*, 83–105.

(346) Wu, Q.; Van Voorhis, T. Direct calculation of electron transfer parameters through constrained density functional theory. *J. Phys. Chem. A* **2006**, *110*, 9212–9218.

(347) Wu, Q.; Van Voorhis, T. Extracting electron transfer coupling elements from constrained density functional theory. *J. Chem. Phys.* **2006**, *125*, 164105.

(348) Van Voorhis, T.; Kowalczyk, T.; Kaduk, B.; Wang, L.-P.; Cheng, C.-L.; Wu, Q. The diabatic picture of electron transfer, reaction barriers, and molecular dynamics. *Annu. Rev. Phys. Chem.* **2010**, *61*, 149–170.

(349) Goldey, M. B.; Brawand, N. P.; Vörös, M.; Galli, G. Charge Transport in Nanostructured Materials: Implementation and Verification of Constrained Density Functional Theory. *J. Chem. Theory Comput.* **2017**, *13*, 2581V2590.

(350) Wu, Q.; Yang, W. A direct optimization method for calculating density functionals and exchange–correlation potentials from electron densities. *J. Chem. Phys.* **2003**, *118*, 2498–2509.

(351) Horn, P. R.; Head-Gordon, M. Alternative definitions of the frozen energy in energy decomposition analysis of density functional theory calculations. *J. Chem. Phys.* **2016**, *144*, 084118.

(352) Mo, Y.; Wang, C.; Guan, L.; Braïda, B.; Hiberty, P. C.; Wu, W. On the nature of blueshifting hydrogen bonds. *Chem.-Eur. J.* **2014**, *20*, 8444–8452.

(353) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(354) Bauschlicher Jr, C. W.; Bagus, P. S.; Nelin, C. J.; Roos, B. O. The nature of the bonding in XCO for X= Fe, Ni, and Cu. *J. Chem. Phys.* **1986**, *85*, 354–364.

(355) Lupinetti, A. J.; Fau, S.; Frenking, G.; Strauss, S. H. Theoretical analysis of the bonding between CO and positively charged atoms. *J. Phys. Chem. A* **1997**, *101*, 9551–9559.

(356) Hobza, P.; Havlas, Z. Blue-shifting hydrogen bonds. *Chem. Rev.* **2000**, *100*, 4253–4264.

(357) Hermansson, K. Blue-shifting hydrogen bonds. *J. Phys. Chem. A* **2002**, *106*, 4695–4702.

(358) Demerdash, O.; Mao, Y.; Liu, T.; Head-Gordon, M.; Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *J. Chem. Phys.* **2017**, *147*, 161721.

(359) Jing, Z.; Qi, R.; Liu, C.; Ren, P. Study of interactions between metal ions and protein model compounds by energy decomposition analyses and the AMOEBA force field. *J. Chem. Phys.* **2017**, *147*, 161733.

(360) Lin, I.-C.; Seitsonen, A. P.; Tavernelli, I.; Rothlisberger, U. Structure and Dynamics of Liquid Water from ab Initio Molecular Dynamics — Comparison of BLYP, PBE, and revPBE Density Functionals with and without van der Waals Corrections. *J. Chem. Theory Comput.* **2012**, *8*, 3902–3910.

(361) Legon, A. C. The halogen bond: an interim perspective. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7736–7747.

(362) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. Evaluation of charge penetration between distributed multipolar expansions. *J. Chem. Phys.* **2000**, *112*, 7300–7306.

(363) Slipchenko, L.; Gordon, M. S. Electrostatic energy in the effective fragment potential method: Theory and application to benzene dimer. *J. Comput. Chem.* **2007**, *28*, 276–291.

(364) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. Improved formulas for the calculation of the electrostatic contribution to the intermolecular interaction energy from multipolar expansion of the electronic distribution. *J. Phys. Chem. A* **2003**, *107*, 10353–10359.

(365) Stone, A. J. Electrostatic damping functions and the penetration energy. *J. Phys. Chem. A* **2011**, *115*, 7017–7027.

(366) Wang, B.; Truhlar, D. G. Screened electrostatic interactions in molecular mechanics. *J. Chem. Theory Comput.* **2014**, *10*, 4480–4487.

(367) Narth, C.; Lagardere, L.; Polack, E.; Gresh, N.; Wang, Q.; Bell, D. R.; Rackers, J. A.; Ponder, J. W.; Ren, P. Y.; Piquemal, J.-P. Scalable improvement of SPME multipolar electrostatics in anisotropic polarizable molecular mechanics using a short-range penetration correction up to quadrupoles. *J. Comput. Chem.* **2016**, *37*, 494–506.

(368) Lee, A. J.; Rick, S. W. The effects of charge transfer on the properties of liquid water. *J. Chem. Phys.* **2011**, *134*, 184507.

(369) Soniat, M.; Rick, S. W. The effects of charge transfer on the aqueous solvation of ions. *J. Chem. Phys.* **2012**, *137*, 044511.

(370) Soniat, M.; Rick, S. W. Charge transfer effects of ions at the liquid water/vapor interface. *J. Chem. Phys.* **2014**, *140*, 184703.

(371) Soniat, M.; Rick, S. W. Charge transfer models of zinc and magnesium in water. *J. Chem. Theory Comput.* **2015**, *11*, 1658–1667.

(372) Soniat, M.; Kumar, R.; Rick, S. W. Hydrated proton and hydroxide charge transfer at the liquid/vapor interface of water. *J. Chem. Phys.* **2015**, *143*, 044702.

(373) Soniat, M.; Pool, G.; Franklin, L.; Rick, S. W. Ion association in aqueous solution. *Fluid Phase Equilibria* **2016**, *407*, 31–38.

(374) Chelli, R.; Pagliai, M.; Procacci, P.; Cardini, G. Polarization response of water and methanol investigated by a polarizable force field and density functional theory calculations: Implications for charge transfer. *J. Chem. Phys.* **2005**, *122*, 074504.

(375) Piquemal, J.-P.; Chevreau, H.; Gresh, N. Toward a separate reproduction of the contributions to the Hartree-Fock and DFT intermolecular interaction energies by polarizable molecular mechanics with the SIBFA potential. *J. Chem. Theory Comput.* **2007**, *3*, 824–837.

(376) Millot, C.; Soetens, J.-C.; Costa, M. T.C. M.; Hodges, M. P.; Stone, A. J. Revised anisotropic site potentials for the water dimer and calculated properties. *J. Phys. Chem. A* **1998**, *102*, 754–770.

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

(378) Chalasinski, G.; Szczesniak, M. M. Origins of structure and energetics of van der Waals clusters from ab initio calculations. *Chem. Rev.* **1994**, *94*, 1723–1765.

(379) Hohenstein, E. G.; Sherrill, C. D. Wavefunction methods for noncovalent interactions. *WIREs Comput. Mol. Sci.* **2012**, *2*, 304–326.

(380) Jorgensen, W. L. Theoretical studies of medium effects on conformational equilibria. *J. Phys. Chem.* **1983**, *87*, 5304–5314.

(381) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free energies of hydration and pure liquid properties of hydrocarbons from the OPLS all-atom model. *J. Phys. Chem.* **1994**, *98*, 13077–13082.

(382) Yin, D.; MacKerell Jr., A. D. Combined ab initio/empirical approach for optimization of Lennard-Jones parameters. *J. Comput. Chem.* **1998**, *19*, 334–348.

(383) MacKerell Jr., A. D. In *Computational Biochemistry and Biophysics*, Becker, O. M., MacKerell Jr., A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker, Inc.: New York, New York, U.S.A., 2001; Chapter 2, pp 7–38.

(384) Ren, P.; Ponder, J. W. Consistent treatment of inter-and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem.* **2002**, *23*, 1497–1506.

(385) Grossfield, A.; Ren, P.; Ponder, J. W. Ion solvation thermodynamics from simulations with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.

(386) Xiang, J. Y.; Ponder, J. W. A valence bond model for aqueous Cu (II) and Zn (II) ions in the AMOEBA polarizable force field. *J. Comput. Chem.* **2013**, *34*, 739–749.

(387) Xiang, J. Y.; Ponder, J. W. An Angular Overlap Model for Cu (II) Ion in the AMOEBA Polarizable Force Field. *J. Chem. Theory Comput.* **2013**, *10*, 298–311.

(388) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Ponder, J. W.; Ren, P. Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046–4063.

(389) Kumar, R.; Wang, F.-F.; Jenness, G. R.; Jordan, K. D. A second generation distributed point polarizable water model. *J. Chem. Phys.* **2010**, *132*, 014309.

(390) Tafipolsky, M.; Engels, B. Accurate intermolecular potentials with physically grounded electrostatics. *J. Chem. Theory Comput.* **2011**, *7*, 1791–1803.

(391) Ansorg, K.; Tafipolsky, M.; Engels, B. Cation- π Interactions: Accurate Intermolecular Potential from Symmetry-Adapted Perturbation Theory. *J. Phys. Chem. B* **2013**, *117*, 10093–10102.

(392) McDaniel, J. G.; Schmidt, J. Physically-motivated force fields from symmetry-adapted perturbation theory. *J. Phys. Chem. A* **2013**, *117*, 2053–2066.

(393) Tafipolsky, M.; Ansorg, K. Toward a Physically Motivated Force Field: Hydrogen Bond Directionality from a Symmetry-Adapted Perturbation Theory Perspective. *J. Chem. Theory Comput.* **2016**, *12*, 1267–1279.

(394) Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. Beyond BornMayer: Improved Models for Short-Range Repulsion in ab Initio Force Fields. *J. Chem. Theory Comput.* **2016**, *12*, 3851–3870.

(395) Misquitta, A. J.; Stone, A. J. Ab initio atom–atom potentials using CamCASP: Theory and application to many-body models for the pyridine dimer. *J. Chem. Theory Comput.* **2016**, *12*, 4184–4208.

(396) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. Towards a force field based on density fitting. *J. Chem. Phys.* **2006**, *124*, 104101.

(397) Zhou, N.; Lu, Z.; Wu, Q.; Zhang, Y. Improved parameterization of interatomic potentials for rare gas dimers with density-based energy decomposition analysis. *J. Chem. Phys.* **2014**, *140*, 214117.

(398) Heßelmann, A.; Jansen, G. First-order intermolecular interaction energies from Kohn–Sham orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

(399) Heßelmann, A.; Jansen, G. Intermolecular induction and exchange-induction energies from coupled-perturbed Kohn–Sham density functional theory. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

(400) Heßelmann, A.; Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

(401) Jansen, G. Symmetry-adapted perturbation theory based on density functional theory for noncovalent interactions. *WIREs Comput. Mol. Sci.* **2014**, *4*, 127–144.

(402) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **2003**, *91*, 033201.

(403) Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.* **2005**, *122*, 214109.

(404) Misquitta, A. J.; Szalewicz, K. Symmetry-adapted perturbation-theory calculations of intermolecular forces employing density-functional description of monomers. *J. Chem. Phys.* **2005**, *122*, 214109.

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

(406) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Density-fitting method in symmetry-adapted perturbation theory based on Kohn-Sham description of monomers. *J. Chem. Theory Comput.* **2006**, *2*, 400–412.

(407) Corminboeuf, C. Minimizing density functional failures for non-covalent interactions beyond van der Waals complexes. *Acc. Chem. Res.* **2014**, *47*, 3217–3224.

(408) DiStasio Jr, R. A.; Gobre, V. V.; Tkatchenko, A. Many-body van der Waals interactions in molecules and condensed matter. *J. Phys.: Condens. Matter* **2014**, *26*, 213202.

(409) Stone, A. J. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chem. Phys. Lett.* **1981**, *83*, 233–239.

(410) Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.* **1981**, *59*, 341–350.

(411) Van Duijnen, P. T.; Swart, M. Molecular and atomic polarizabilities: Thole's model revisited. *J. Phys. Chem. A* **1998**, *102*, 2399–2407.

(412) Halgren, T. A. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.

(413) Ponder, J. W., *Tinker–Software Tools for Molecular Design*, Tinker 7.1; Washington University: St. Louis, MO, 2015.

(414) Wu, J. C.; Piquemal, J.-P.; Chaudret, R.; Reinhardt, P.; Ren, P. Polarizable molecular dynamics simulation of Zn(II) in water using the AMOEBA force field. *J. Chem. Theory Comput.* **2010**, *6*, 2059–2070.

(415) Woon, D. E.; Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *J. Chem. Phys.* **1995**, *103*, 4572–4585.

(416) Peterson, K. A.; Dunning Jr, T. H. Accurate correlation consistent basis sets for molecular core–valence correlation effects: The second row atoms Al–Ar, and the first row atoms B–Ne revisited. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

(417) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N_2, and H_2O. *Chem. Phys. Lett.* **1998**, *286*, 243–252.

(418) Gresh, N.; Garmer, D. R. Comparative binding energetics of Mg^{2+}, Ca^{2+}, Zn^{2+}, and Cd^{2+} to biologically relevant ligands: Combined ab initio SCF supermolecule and molecular mechanics investigation. *J. Comput. Chem.* **1996**, *17*, 1481–1495.

(419) Gresh, N. Model, multiply hydrogen-bonded water oligomers (N = 3–20). How closely can a separable, ab initio-grounded molecular mechanics procedure reproduce the results of supermolecule quantum chemical computations? *J. Phys. Chem. A* **1997**, *101*, 8680–8694.

(420) Giese, T. J.; York, D. M. Many-body force field models based solely on pairwise Coulomb screening do not simultaneously reproduce correct gas-phase and condensed-phase polarizability limits. *J. Chem. Phys.* **2004**, *120*, 9903–9906.

(421) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. Key role of the polarization anisotropy of water in modeling classical polarizable force fields. *J. Phys. Chem. A* **2007**, *111*, 8170–8176.

(422) Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; MacKerrel Jr., A. D.; Roux, B. Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field based on the classical Drude oscillator. *J. Chem. Theory Comput.* **2006**, *2*, 1587–1597.

(423) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **2005**, *105*, 2999–3093.

(424) Born, M Volumes and hydration warmth of ions. *Z. Phys.* **1920**, *1*, 45–48.

(425) Onsager, L. Electric moments of molecules in liquids. *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.

(426) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Self-Consistent Reaction Field Model for Aqueous and Nonaqueous Solutions Based on Accurate Polarized Partial Charges. *J. Chem. Theory Comput.* **2007**, 2011–2033.

(427) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Performance of SM8 on a test to predict small-molecule solvation free energies. *J. Phys. Chem. B* **2008**, *112*, 8651–8655.

(428) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized Born Solvation Model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609–620.

(429) Klamt, A; Schüürmann, G COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans. 2* **1993**, *1993*, 799–805.

(430) Truong, T. N.; Stefanovich, E. V. A new method for incorporating solvent effect into the classical, ab initio molecular orbital and density functional theory frameworks for arbitrary shape cavity. *Chem. Phys. Lett.* **1995**, *240*, 253–260.

(431) Mennucci, B.; Cancès, E.; Tomasi, J. Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications. *J. Phys. Chem. B* **1997**, *101*, 10506–10517.

(432) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032.

(433) Truong, T. N. Solvent Effects on Structure and Reaction Mechanism: A Theoretical Study of [2 + 2] Polar Cycloaddition between Ketene and Imine. *J. Phys. Chem. B* **1998**, *102*, 7877–7881.

(434) Tomasi, J; Mennucci, B; Cancès, E The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. *J. Mol. Struct. THEOCHEM* **1999**, *464*, 211–226.

(435) Chipman, D. M. Reaction field treatment of charge penetration. *J. Chem. Phys.* **2000**, *112*, 5558.

(436) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(437) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(438) Lange, A. W.; Herbert, J. M. Polarizable Continuum Reaction-Field Solvation Models Affording Smooth Potential Energy Surfaces. *J. Phys. Chem. Lett.* **2010**, *1*, 556–561.

(439) Lange, A. W.; Herbert, J. M. A smooth, nonsingular, and faithful discretization scheme for polarizable continuum models: the switching/Gaussian approach. *J. Chem. Phys.* **2010**, *133*, 244111.

(440) You, Z.-Q.; Herbert, J. M. Reparameterization of an Accurate, Few-Parameter Implicit Solvation Model for Quantum Chemistry: Composite Method for Implicit Representation of Solvent, CMIRS v. 1.1. *J. Chem. Theory Comput.* **2016**, *12*, 4338–4346.

(441) Fattebert, J.-L.; Gygi, F. Density functional theory for efficient ab initio molecular dynamics simulations in solution. *J. Comput. Chem.* **2002**, *23*, 662–666.

(442) Fattebert, J.-L.; Gygi, F. First-principles molecular dynamics simulations in a continuum solvent. *Int. J. Quant. Chem.* **2003**, *93*, 139–147.

(443) Scherlis, D. A.; Fattebert, J.-L.; Gygi, F.; Cococcioni, M.; Marzari, N. A unified electrostatic and cavitation model for first-principles molecular dynamics in solution. *J. Chem. Phys.* **2006**, *124*, 074103.

(444) Dziedzic, J.; Helal, H. H.; Skylaris, C.-K.; Mostofi, A. A.; Payne, M. C. Minimal parameter implicit solvent model for ab initio electronic-structure calculations. *Europhys. Lett.* **2011**, *95*, 43001.

(445) Cramer, C. J.; Truhlar, D. G. A universal approach to solvation modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(446) Klamt, A; Mennucci, B; Tomasi, J; Barone, V; Curutchet, C; Orozco, M; Luque, F. J. On the Performance of Continuum Solvation Methods. A Comment on Universal Approaches to Sovation Modeling". *Acc. Chem. Res.* **2009**, *42*, 489–492.

(447) Laasonen, K.; Sprik, M.; Parrinello, M.; Car, R. "Ab initio" liquid water. *J. Chem. Phys.* **1993**, *99*, 9080.

(448) Sprik, M.; Hutter, J.; Parrinello, M. Ab initio molecular dynamics simulation of liquid water: comparison of three gradient-corrected density functionals. *J. Chem. Phys.* **1996**, *105*, 1142–1152.

(449) Silvestrelli, P. L.; Parrinello, M. Structural, Electronic, and Bonding Properties of Liquid Water from First Principles. *J. Chem. Phys.* **1999**, *111*, 3572–3580.

(450) McGrath, M. J.; Siepmann, J. I.; Kuo, I. F. W.; Mundy, C. J.; Vandevondele, J.; Hutter, J.; Mohamed, F.; Krack, M. Isobaric-isothermal monte carlo simulations from first principles: Application to liquid water at ambient conditions. *ChemPhysChem* **2005**, *6*, 1894–1901.

(451) Zhang, C.; Donadio, D.; Gygi, F.; Galli, G. First principles simulations of the infrared spectrum of liquid water using hybrid density functionals. *J. Chem. Theory Comput.* **2011**, *7*, 1443–1449.

(452) Del Ben, M.; Schönherr, M.; Hutter, J.; VandeVondele, J. Bulk Liquid Water at Ambient Temperature and Pressure from MP2 Theory. *J. Phys. Chem. Lett.* **2013**, *4*, 3753–3759.

(453) Gaiduk, A. P.; Gygi, F.; Galli, G. Density and compressibility of liquid water and ice from first-principles simulations with hybrid functionals. *J. Phys. Chem. Lett.* **2015**, *6*, 2902–2908.

(454) White, J. A.; Schwegler, E.; Galli, G.; Gygi, F. The solvation of Na^+ in water: First-principles simulations. *J. Chem. Phys.* **2000**, *113*, 4668–4673.

(455) Lightstone, F. C.; Schwegler, E.; Hood, R. Q.; Gygi, F.; Galli, G. A first principles molecular dynamics simulation of the hydrated magnesium ion. *Chem. Phys. Lett.* **2001**, *343*, 549–555.

(456) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. A First-Principles Molecular Dynamics Study of Calcium in Water. *ChemPhysChem* **2005**, *6*, 1745–1749.

(457) Ikeda, T.; Boero, M.; Terakura, K. Hydration of alkali ions from first principles molecular dynamics revisited. *J. Chem. Phys.* **2007**, *126*, 01B611.

(458) Schmidt, D. A.; Scipioni, R.; Boero, M. Water Solvation Properties: An Experimental and Theoretical Investigation of Salt Solutions at Finite Dilution. *J. Phys. Chem. A* **2009**, *113*, 7725–7729.

(459) Zhang, C.; Pham, T. A.; Gygi, F.; Galli, G. Communication: Electronic structure of the solvated chloride anion from first principles molecular dynamics. *J. Chem. Phys.* **2013**, *138*, 181102.

(460) Gaiduk, A. P.; Zhang, C.; Gygi, F.; Galli, G. Structural and electronic properties of aqueous NaCl solutions from ab initio molecular dynamics simulations with hybrid density functionals. *Chem. Phys. Lett.* **2014**, *604*, 89–96.

(461) Gaiduk, A. P.; Galli, G. Local and Global Effects of Dissolved Sodium Chloride on the Structure of Water. *J. Phys. Chem. Lett.* **2017**, *8*, 1496–1502.

(462) Schmidt, J.; Vandevondele, J.; William Kuo, I. F.; Sebastiani, D.; Ilja Siepmann, J.; Hutter, J.; Mundy, C. J. Isobaric-isothermal molecular dynamics simulations utilizing density functional theory: An assessment of the structure and density of water at near-ambient conditions. *J. Phys. Chem. B* **2009**, *113*, 11959–11964.

(463) Wang, J.; Roman-Perez, G.; Soler, J. M.; Artacho, E.; Fernandez-Serra, M.-V. Density, structure, and dynamics of water: The effect of van der Waals interactions. *J. Chem. Phys.* **2011**, *134*, 024516.

(464) Ma, Z.; Zhang, Y.; Tuckerman, M. E. Ab initio molecular dynamics study of water at constant pressure using converged basis sets and empirical dispersion corrections. *J. Chem. Phys.* **2012**, *137*, 044506.

(465) Forster-Tonigold, K.; Groß, A. Dispersion corrected RPBE studies of liquid water. *J. Chem. Phys.* **2014**, *141*, 064501.

(466) DiStasio, R. A.; Santra, B.; Li, Z.; Wu, X.; Car, R. The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water. *J. Chem. Phys.* **2014**, *141*, 084502.

(467) Bankura, A.; Santra, B.; DiStasio, R. A.; Swartz, C. W.; Klein, M. L.; Wu, X. A systematic study of chloride ion solvation in water using van der Waals inclusive hybrid density functional theory. *Mol. Phys.* **2015**, *113*, 2842–2854.

(468) Gillan, M. J.; Alfè, D.; Michaelides, A. Perspective: How good is DFT for water? *J. Chem. Phys.* **2016**, *144*, 130901.

(469) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(470) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(471) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(472) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(473) Gurunathan, P. K.; Acharya, A.; Ghosh, D.; Kosenkov, D.; Kaliman, I.; Shao, Y.; Krylov, A. I.; Slipchenko, L. V. Extension of the Effective Fragment Potential Method to Macromolecules. *J. Phys. Chem. B* **2016**, *120*, 6562–6574.

(474) Olsen, J. M.; Aidas, K.; Kongsted, J. Excited states in solution through polarizable embedding. *J. Chem. Theory Comput.* **2010**, *6*, 3721–3734.

(475) Steindal, A. H.; Ruud, K.; Frediani, L.; Aidas, K.; Kongsted, J. Excitation energies in solution: the fully polarizable QM/MM/PCM method. *J. Phys. Chem. B* **2011**, *115*, 3027–3037.

(476) Schwabe, T.; Olsen, J. M. H.; Sneskov, K.; Kongsted, J.; Christiansen, O. Solvation effects on electronic transitions: Exploring the performance of advanced solvent potentials in polarizable embedding calculations. *J. Chem. Theory Comput.* **2011**, *7*, 2209–2217.

(477) Kratz, E. G.; Walker, A. R.; Lagardère, L.; Lipparini, F.; Piquemal, J. P.; Cisneros, G. A. LICHEM: A QM/MM program for simulations with multipolar and polarizable force fields. *J. Comput. Chem.* **2016**, *37*, 1019–1029.

(478) Loco, D.; Polack, E.; Caprasecca, S.; Lagardère, L.; Lipparini, F.; Piquemal, J.-P.; Mennucci, B. A QM/MM Approach Using the AMOEBA Polarizable Embedding: From Ground State Energies to Electronic Excitations. *J. Chem. Theory Comput.* **2016**, *12*, 3654–3661.

(479) Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method. *J. Chem. Phys.* **2002**, *117*, 10534–10547.

(480) Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A. Quantum mechanics/molecular mechanics electrostatic embedding with continuous and discrete functions. *J. Chem. Phys. B* **2006**, *110*, 13682–13684.

(481) Cisneros, G. A.; Tholander, S; Parisel, O; Darden, T.; Elking, D; Perera, L; Piquemal, J.-P. Simple formulas for improved point-charge electrostatics in classical force fields and hybrid quantum mechanical/molecular mechanical embedding. *Int. J. Quant. Chem.* **2008**, *108*, 1905–1912.

(482) Biswas, P.; Gogonea, V. A polarizable force-field model for quantum-mechanical-molecular-mechanical Hamiltonian using expansion of point charges into orbitals. *J. Chem. Phys.* **2008**, *129*, 154108.

(483) Wang, B.; Truhlar, D. G. Including charge penetration effects in molecular modeling. *J. Chem. Theory Comput.* **2010**, *6*, 3330–3342.

(484) Wang, B.; Truhlar, D. G. Partial atomic charges and screened charge models of the electrostatic potential. *J. Chem. Theory Comput.* **2012**, *8*, 1989–1998.

(485) Woodcock, H. L.; Hodošček, M.; Gilbert, A. T.; Gill, P. M.; Schaefer, H. F.; Brooks, B. R. Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations. *J. Comput. Chem.* **2007**, *28*, 1485–1502.

(486) Lu, Z.; Zhang, Y. Interfacing ab initio quantum mechanical method with classical Drude osillator polarizable model for molecular dynamics simulation of chemical reactions. *J. Chem. Theory Comput.* **2008**, *4*, 1237–1248.

(487) Boulanger, E.; Thiel, W. Solvent boundary potentials for hybrid QM/MM computations using classical drude oscillators: a fully polarizable model. *J. Chem. Theory Comput.* **2012**, *8*, 4527–4538.

(488) Boulanger, E.; Thiel, W. Toward QM/MM simulation of enzymatic reactions with the drude oscillator polarizable force field. *J. Chem. Theory Comput.* **2014**, *10*, 1795–1809.

(489) Bryce, R. A.; Buesnel, R.; Hillier, I. H.; Burton, N. A. A solvation model using a hybrid quantum mechanical/molecular mechanical potential with fluctuating solvent charges. *Chem. Phys. Lett.* **1997**, *279*, 367–371.

(490) Lipparini, F.; Cappelli, C.; Barone, V. Linear response theory and electronic transition energies for a fully polarizable QM/classical Hamiltonian. *J. Chem. Theory Comput.* **2012**, *8*, 4153–4165.

(491) Lipparini, F.; Cappelli, C.; Barone, V. A gauge invariant multiscale approach to magnetic spectroscopies in condensed phase: General three-layer model, computational implementation and pilot applications. *J. Chem. Phys.* **2013**, *138*, 234108.

(492) Thompson, M. A.; Schenter, G. K. Excited states of the bacteriochlorophyll b dimer of Rhodopseudomonas viridis: a QM/MM study of the photosynthetic reaction center that includes MM polarization. *J. Chem. Phys.* **1995**, *99*, 6374–6386.

(493) Aida, M; Yamataka, H; Dupuis, M. Critical assessment of the hybrid QM/MM-polvib approach: Small water clusters using polarizable flexible water potentials. *Int. J. Quant. Chem.* **2000**, *77*, 199–210.

(494) Nielsen, C. B.; Christiansen, O.; Mikkelsen, K. V.; Kongsted, J. Density functional self-consistent quantum mechanics/molecular mechanics theory for linear and nonlinear molecular properties: Applications to solvated water and formaldehyde. *J. Chem. Phys.* **2007**, *126*, 154112.

(495) Olsen, J. M. H.; Steinmann, C.; Ruud, K.; Kongsted, J. Polarizable Density Embedding: A New QM/QM/MM-Based Computational Strategy. *J. Phys. Chem. A* **2015**, *119*, 5344–5355.

(496) Curutchet, C.; Muñoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. Electronic energy transfer in condensed phase studied by a polarizable QM/MM model. *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848.

(497) Thellamurege, N. M.; Si, D.; Cui, F.; Zhu, H.; Lai, R.; Li, H. QuanPol: A full spectrum and seamless QM/MM program. *J. Comput. Chem.* **2013**, *34*, 2816–2833.

(498) Schwörer, M.; Breitenfeld, B.; Tröster, P.; Bauer, S.; Lorenzen, K.; Tavan, P.; Mathias, G. Coupling density functional theory to polarizable force fields for efficient and accurate Hamiltonian molecular dynamics simulations. *J. Chem. Phys.* **2013**, *138*, 244103.

(499) Schwörer, M.; Wichmann, C.; Tavan, P. A polarizable QM/MM approach to the molecular dynamics of amide groups solvated in water. *J. Chem. Phys.* **2016**, *144*, 114504.

(500) Caprasecca, S.; Curutchet, C.; Mennucci, B. Toward a Unified Modeling of Environment and Bridge-Mediated Contributions to Electronic Energy Transfer: A Fully Polarizable QM/MM/PCM Approach. *J. Chem. Theory Comput.* **2012**, *8*, 4462–4473.

(501) Caprasecca, S.; Jurinovich, S.; Viani, L.; Curutchet, C.; Mennucci, B. Geometry optimization in polarizable QM/MM models: the induced dipole formulation. *J. Chem. Theory Comput.* **2014**, *10*, 1588–1598.

(502) Caprasecca, S.; Jurinovich, S.; Lagardere, L.; Stamm, B.; Lipparini, F. Achieving Linear Scaling in Computational Cost for a Fully Polarizable MM/Continuum Embedding. *J. Chem. Theory Comput.* **2015**, *11*, 694–704.

(503) Shaw, K. E.; Woods, C. J.; Mulholland, A. J. Compatibility of Quantum Chemical Methods and Empirical (MM) Water Models in Quantum Mechanics/Molecular Mechanics Liquid Water Simulations. *J. Phys. Chem. Lett.* **2010**, *1*, 219–223.

(504) König, G.; Mei, Y.; Pickard, F. C.; Simmonett, A. C.; Miller, B. T.; Herbert, J. M.; Woodcock, H. L.; Brooks, B. R.; Shao, Y. Computation of Hydration Free Energies Using the Multiple Environment Single System Quantum Mechanical/Molecular Mechanical (MESS-QM/MM) Method. *J. Chem. Theory Comput.* **2016**, *12*, 332–344.

(505) Pickard, F. C.; König, G.; Simmonett, A. C.; Shao, Y.; Brooks, B. R. An efficient protocol for obtaining accurate hydration free energies using quantum chemistry and reweighting from molecular dynamics simulations. *Bioorg. Med. Chem.* **2016**, *24*, 4988–4997.

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

(507) Kaliman, I. A.; Slipchenko, L. V. LIBEFP: A new parallel implementation of the effective fragment potential method as a portable software library. *J. Comput. Chem.* **2013**, *34*, 2284–2292.

(508) Kaliman, I. A.; Slipchenko, L. V. Hybrid MPI/OpenMP parallelization of the effective fragment potential method in the libefp software library. *J. Comput. Chem.* **2015**, *36*, 129–135.

(509) Gao, J. Energy components of aqueous solution: insight from hybrid QM/MM simulations using a polarizable solvent model. *J. Comput. Chem.* **1997**, *18*, 1061–1071.

(510) Thellamurege, N. M.; Hirao, H. Effect of protein environment within cytochrome P450cam evaluated using a polarizable-embedding QM/MM method. *J. Phys. Chem. B* **2014**, *118*, 2084–2092.

(511) Gordon, M. S.; Slipchenko, L.; Li, H.; Jensen, J. H., *Annual reports in computational chemistry*; Elsevier: 2007; Vol. 3, pp 177–193.

(512) Grossfield, A. Dependence of ion hydration on the sign of the ions charge. *J. Chem. Phys.* **2005**, *122*, 024506.

(513) Deible, M. J.; Tuguldur, O.; Jordan, K. D. Theoretical study of the binding energy of a methane molecule in a $(H_2O)_{20}$ dodecahedral cage. *J. Phys. Chem. B* **2014**, *118*, 8257–8263.

(514) Fox, S. J.; Pittock, C.; Fox, T.; Tautermann, C. S.; Malcolm, N.; Skylaris, C.-K. Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules. *J. Chem. Phys.* **2011**, *135*, 224107.

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

(516) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J. How large should the QM region be in QM/MM calculations? The case of catechol O-methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381.

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

(518) Bulo, R. E.; Ensing, B.; Sikkema, J.; Visscher, L. Toward a practical method for adaptive QM/MM simulations. *J. Chem. Theory Comput.* **2009**, *5*, 2212–2221.

(519) Nielsen, S. O.; Bulo, R. E.; Moore, P. B.; Ensing, B. Recent progress in adaptive multiscale molecular dynamics simulations of soft matter. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12401–12414.

(520) Wheatley, R. J. Gaussian multipole functions for describing molecular charge distributions. *Mol. Phys.* **1993**, *79*, 597–610.

(521) Wheatley, R. J.; Mitchell, J. B. Gaussian multipoles in practice: Electrostatic energies for intermolecular potentials. *J. Comput. Chem.* **1994**, *15*, 1187–1198.

(522) Elking, D. M.; Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A.; Pedersen, L. G. Gaussian multipole model (GMM). *J. Chem. Theory Comput.* **2010**, *6*, 190–202.

(523) Cisneros, G. A. Application of gaussian electrostatic model (GEM) distributed multipoles in the AMOEBA force field. *J. Chem. Theory Comput.* **2012**, *8*, 5072–5080.

(524) Spackman, M. A. The use of the promolecular charge density to approximate the penetration contribution to intermolecular electrostatic energies. *Chem. Phys. Lett.* **2006**, *418*, 158–162.

(525) Albaugh, A. et al. Advanced Potential Energy Surfaces for Molecular Simulation. *J. Phys. Chem. B* **2016**, *120*, 9811–9832.

(526) Nâbo, L. J.; Olsen, J. M. H.; Holmgaard List, N.; Solanko, L. M.; Wüstner, D.; Kongsted, J. Embedding beyond electrostatics—The role of wave function confinement. *J. Chem. Phys.* **2016**, *145*.

(527) Fradelos, G.; Wesołowski, T. A. Importance of the Intermolecular Pauli Repulsion in Embedding Calculations for Molecular Properties: The Case of Excitation Energies for a Chromophore in Hydrogen-Bonded Environments. *J. Phys. Chem. A* **2011**, *115*, 10018–10026.

(528) Vaidehi, N.; Wesolowski, T. A.; Warshel, A. Quantum-mechanical calculations of solvation free energies. A combined abinitio pseudopotential free-energy perturbation approach. *J. Chem. Phys.* **1992**, *97*, 4264–4271.

(529) Chalmet, S.; Ruiz-López, M. F. New approaches to the description of short-range repulsion interactions in hybrid quantum/classical systems. *Chem. Phys. Lett.* **2000**, *329*, 154–159.

(530) Sebastiani, D.; Rothlisberger, U. Nuclear magnetic resonance chemical shifts from hybrid DFT QM/MM calculations. *J. Phys. Chem. B* **2004**, *108*, 2807–2815.

(531) Jin, Y.; Johnson, E. R.; Hu, X.; Yang, W.; Hu, H. Contributions of pauli repulsions to the energetics and physical properties computed in QM/MM methods. *J. Comput. Chem.* **2013**, *34*, 2380–2388.

(532) Wesolowski, T. A.; Warshel, A. Frozen density functional approach for ab initio calculations of solvated molecules. *J. Phys. Chem.* **1993**, *97*, 8050–8053.

(533) Wesolowski, T. A.; Weber, J. Kohn-Sham equations with constrained electron density: an iterative evaluation of the ground-state electron density of interacting molecules. *Chem. Phys. Lett.* **1996**, *248*, 71–76.

(534) Wesolowski, T. A., *Computational chemistry: reviews of current trends*; World Scientific: Singapore: 2006; Vol. 10, pp 1–82.

(535) Barandiarán, Z.; Seijo, L. The ab initio model potential representation of the crystalline environment. Theoretical study of the local distortion on NaCl:Cu^+. *J. Chem. Phys.* **1988**, *89*, 5739–5746.

(536) Swerts, B.; Chibotaru, L. F.; Lindh, R.; Seijo, L.; Barandiaran, Z.; Clima, S.; Pierloot, K.; Hendrickx, M. F. Embedding fragment ab initio model potentials in CASSCF/CASPT2 calculations of doped solids: implementation and applications. *J. Chem. Theory Comput.* **2008**, *4*, 586–594.

(537) Ben-Nun, M; Martinez, T. J. Direct evaluation of the Pauli repulsion energy using 'classical' wavefunctions in hybrid quantum/classical potential energy surfaces. *Chem. Phys. Lett.* **1998**, *290*, 289–295.

(538) Öhrn, A.; Karlström, G. A theoretical study of the solvent shift to the transition in formaldehyde with an effective discrete quantum chemical solvent model including non-electrostatic perturbation. *Mol. Phys.* **2006**, *104*, 3087–3099.

(539) Söderhjelm, P.; Öhrn, A. On the coupling of intermolecular polarization and repulsion through pseudo-potentials. *Chem. Phys. Lett.* **2009**, *468*, 94–99.

(540) Gresh, N.; Claverie, P.; Pullman, A. Intermolecular interactions: Elaboration on an additive procedure including an explicit charge-transfer contribution. *Int. J. Quant. Chem.* **1986**, *29*, 101–118.

(541) Gresh, N.; Piquemal, J.-P.; Krauss, M. Representation of Zn (II) complexes in polarizable molecular mechanics. Further refinements of the electrostatic and short-range contributions. Comparisons with parallel ab initio computations. *J. Comput. Chem.* **2005**, *26*, 1113–1130.

(542) Wheatley, R. J.; Price, S. L. An overlap model for estimating the anisotropy of repulsion. *Mol. Phys.* **1990**, *69*, 507–533.

(543) Jensen, J. H.; Gordon, M. S. An approximate formula for the intermolecular Pauli repulsion between closed shell molecules. *Mol. Phys.* **1996**, *89*, 1313–1325.

(544) Jensen, J. H.; Gordon, M. S. An approximate formula for the intermolecular Pauli repulsion between closed shell molecules. II. Application to the effective fragment potential method. *J. Chem. Phys.* **1998**, *108*, 4772.

(545) Gavezzotti, A Calculation of intermolecular interaction energies by direct numerical integration over electron densities. 2. An improved polarization model and the evaluation of dispersion and repulsion energies. *J. Phys. Chem. B* **2003**, *107*, 2344–2353.

(546) Söderhjelm, P.; Karlström, G.; Ryde, U. Comparison of overlap-based models for approximating the exchange-repulsion energy. *J. Chem. Phys.* **2006**, *124*, 244101.

(547) Giese, T. J.; York, D. M. Charge-dependent model for many-body polarization, exchange, and dispersion interactions in hybrid quantum mechanical/molecular mechanical calculations. *J. Chem. Phys.* **2007**, *127*, 194101.

(548) Kuechler, E. R.; Giese, T. J.; York, D. M. Charge-dependent many-body exchange and dispersion interactions in combined QM/MM simulations. *J. Chem. Phys.* **2015**, *143*, 234111.

(549) Kaduk, B.; Tsuchimochi, T.; Van Voorhis, T. Analytic energy gradients for constrained DFT-configuration interaction. *J. Chem. Phys.* **2014**, *140*, 18A503.