Threading a path to exascale with chemical scissors and integral compressors in a singular manner

by

Buu Q. Pham

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Physical Chemistry

Program of Study Committee:
Mark S. Gordon, Major Professor
James W. Evans
Igor I. Slowing
Xueyu Song
Theresa L. Windus

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

DEDICATION

This work is dedicated to my sons, bé Ổi and bé Đậu.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Prof. Mark S. Gordon, whose expertise, understanding, generous guidance and support have made it possible for me to work on research that was of great interest to me. It has been a privilege to work with him.

I would like to thank my program of study committee members, Profs. Theresa L. Windus, Jame W. Evans, Xueyu Song and Igor I. Slowing and Mark S. Gordon for their guidance and support throughout the course of this research.

I would also like to offer my appreciation to doctors Mike Schmidt, Colleen Bertoni, Kris Keipert, Luke Roskop, Sarom Leang; and my friend Viet Q. Tran for walking me through joyful quantum chemistry, computer science and mathematical models.

I would also like to thank all members of Dr. Gordon's group who make me feel I am home.

ABSTRACT

Research presented in this dissertation aims at enabling (correlated) fragmentation methods to explore biochemistry and catalysis effects of macrosystems at high levels of accuracy using exascale computing resources. The target is the second-order MollerPlesset perturbation theory (MP2), and MP2 in the FMO framework (FMO/MP2). First, the 2-electron integral bottleneck is addressed by using the resolution-of-the-identity (RI) approximation to reduce the memory storage and the computational cost of the integral transformation from the atomic orbital (AO) to the molecular orbital (MO) basis. The RI approximation is also combined with the singular value decomposition (SVD) to introduce a flexible compression factor that fully controls the accuracy of the integral compression. The RIMP2 energy and analytic energy gradient are implemented in the GAMESS electronic structure program and are parallelized with an efficient hybrid distributed/shared memory model with the support of the MPI and OpenMP APIs. Both the RI-MP2 energy and gradient are interfaced to the FMO framework for large system calculations.

CHAPTER 1. INTRODUCTION

1.1 Studying objects

Quantum mechanics (QM) can accurately predict molecular properties. However, except for very simple systems (e.g., particle in a box, harmonic oscillator, hydrogen atom), the Schrodinger equations of most molecular systems are solved with approximations and numerical tools. The computational costs grow rapidly with the problem size. In addition to the requirement of large floating-point operations, large *ab initio* problems also encounter high memory demands and communication overhead that introduce a degree of difficulty for efficient parallel code implementation. Systems of more than a hundred heavy atoms are, therefore, usually out of the reach of first principles methods.

Recently developed fragmentation methods, particularly the fragment molecular orbital (FMO) methods,[1–5] are feasible approaches to treat large molecular systems at the accuracy of the underlying (*ab initio*) methods. By taking advantage of the locality of macrosystems, fragmentation methods can (intuitively) partition large systems into small fragments, which can be processed essentially independently. The fragmentation can, in principle, eliminate a large part of redundant 2-electron integrals, reduce the dimension of matrix processing (e.g., matrix diagonalization) and enhance convergence of iterative equation solvers. The fragmentation methods also naturally facilitate parallel code implementation.[6,7] Nevertheless, the computational cost of fragmentation methods with (e.g., dynamic) correlation effects included remains expensive; and the current FMO parallel code implementation is based on an inefficient distributed memory model.[6,7]

Our research presented in this dissertation, therefore, aims at enabling (correlated) fragmentation methods to explore biochemistry and catalysis effects of macrosystems at high levels of accuracy using exascale computing resources. The target is the second-order Moller-Plesset perturbation theory (MP2), and MP2 in the FMO framework (FMO/MP2). First, the 2-electron integral bottleneck is addressed by using the resolution-of-the-identity (RI) approximation to reduce the memory storage and the computational cost of the integral transformation from the atomic orbital (AO) to the molecular orbital (MO) basis. The RI approximation is also combined with the singular value decomposition (SVD) to introduce a flexible compression factor that fully controls the accuracy of the integral compression.[8] The RI-MP2 energy and analytic energy gradient are implemented in the GAMESS electronic structure program[9] and are parallelized with an efficient hybrid distributed/shared memory model with the support of the MPI and OpenMP APIs. Both the RI-MP2 energy and gradient are interfaced to the FMO framework for large system calculations.

The next step of the study is to explore physical and chemical properties of practical macrosystems, particularly heterogeneous catalysis based on mesoporous silica nanoparticles.[10] This includes optimizing the threaded FMO/RI-MP2 codes, and interfacing RI-MP2 to the advanced EFMO[11,12] framework. The heavy computational demand will be processed by accelerators.

1.2. Dissertation organization

This dissertation is organized as follows.

i) Chapter 1 introduces the general theory used in the later chapters.

ii) Chapter 2 presents the integral compressors including the RI and SVD-RI approximations applied to the MP2 correlation energy.

iii) Chapter 3 presents a prototype study combining the FMO method, the RI approximation and the hybrid distributed/shared memory model for the MP2 correlation energy.

iv) Chapter 4 discusses the FMO/RI-MP2 analytic gradient implementation in GAMESS using the hybrid parallel model.

v) Chapters 5 and 6 are applications of *ab initio* methods for graphene and a viewpoint on interpreting molecular properties in terms of molecular orbital concepts.

vi) Chapter 7 is a brief conclusion and outlook for future development.

1.3. Theoretical background

1.3.1. Energy quantization and wave mechanics

The ultraviolet catastrophe in the black body radiation problem led to Max Planck's postulation in 1900 that the energy of oscillators is quantized. Five years later, by adapting the energy quantization idea, Albert Einstein treated light as a bundle of energy quanta (also called photons) and successfully explained the photoelectric effect. Since then light has been considered to exhibit both wave-like and particle-like character, a concept called wave-particle duality, which was extended to matter by Louis de Broglie in 1923. According to de Broglie, the motion of any particle is associated with a wavelength. While the theoretical wavelengths associated with the motion of most objects are negligible compared with their dimension, the wavelengths associated with micro-particles (e.g., electrons and nuclei in atoms and molecules) are relevant. The Germer-Davisson electron diffraction experiments [1923-1927] finally

confirmed de Broglie's audacious conjecture. The motion of electrons and nuclei in atoms and/or molecules must be described by a wave function of space and time $\vec{\Psi}(q,t)$, whose evolution follows the time-dependent equation[13] postulated by Erwin Schrodinger in 1925

$$-i\hbar\frac{\partial}{\partial t}\vec{\Psi}(q,t)=\widehat{H}\vec{\Psi}(q,t) \tag{1}$$

$$\widehat{H}=\hat{T}_e+\hat{T}_N+\hat{V}_{NN}+\hat{V}_{eN}+\hat{V}_{ee}+\hat{V}_{ext} \tag{2}$$

In equation (1), i is the imaginary unit, q is a general spin-space spatial coordinate of all particles, and t is the time variable. $\widehat{H}$ is the total energy (Hamiltonian) operator consisting of the kinetic energy operators describing motion of electrons $\hat{T}_e$ and nuclei $\hat{T}_N$; and the potential energy operators due to the nuclear-nuclear $\hat{V}_{NN}$, electron-nuclear $\hat{V}_{eN}$, and electron-electron $\hat{V}_{ee}$ interactions. When the system is placed in a [static or time-dependent] force field, there can be additional terms to describe the external field $\hat{V}_{ext}$. This dissertation only addresses molecules in vacuum with no external field. Eq. (1) can be solved by factoring the time-dependent wave function into the spin-space $\Psi(\mathrm{q})$ and the time $\psi(\mathrm{t})$ functions that obey the equations

$$\widehat{H}\Psi(q)=E\Psi(q) \tag{3}$$

$$\frac{\partial}{\partial t}\psi(t)=-\frac{i}{\hbar}E\psi(t) \tag{4}$$

Eq. (3) is called the time-*in*dependent Schrodinger equation, in which E is interpreted as the total energy of the system, and $\Psi(\mathrm{q})$ corresponds to the amplitude of a classical wave. This eigenvalue equation is thus also called the amplitude equation. However, unlike classical waves, due to the Heisenberg uncertainty principle,[14] $\Psi(\mathrm{q})$ does not represent the trajectory of

electrons and/or nuclei. The most popular interpretation (called the Born interpretation)[15] of $\Psi(\mathrm{q})$ is that its module $\Psi^*(\mathrm{q})\Psi(\mathrm{q})dq$ is the probability of finding a particle in the volume dq. Since one can always find a particle when searching for the whole space, the wave function is normalized.

$$\int_{-\infty}^{\infty} dq \Psi^*(q)\Psi(q) = 1 \tag{5}$$

In bound states of atoms and molecules, electrons and nuclei are trapped in their electrostatic potential. The normalization requires the corresponding wave functions to vanish at large distances. Such boundary conditions restrict the motions of particles inside atoms or molecules similar to a classical wave with fixed ends or surfaces that yield standing waves with only an integer number of wavelengths allowed. In a similar manner, in their bound states, atoms and molecules can only exist in discrete energy states called quantum states. This does not happen in scattering states when wave functions only need to be normalized. The wave treatment of electrons and nuclei in atoms and molecules, combined with the boundary conditions lead to the quantization of energy observed in classic experiments.

1.3.2 Solving the amplitude equation

Solving the amplitude equation for the spatial wave function $\Psi(\mathrm{q})$ and the total energy is the central problem of quantum chemistry. Dirac noted[16] "The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved". Except for the simplest problems (e.g., particle in a box,

harmonic oscillator, hydrogen atom), the rest of the amplitude equations applied in chemistry are solved with approximations and numerical tools.

An important approximation, the Born-Oppenheimer approximation,[17] decouples the motion of heavy nuclei from electron motions. The electronic state of a molecule is represented by the electronic wave function $\Psi_{elec}(q_{elec})$ that follows the electronic Schrodinger equation.

$$\widehat{H}_{elec}\Psi_{elec}\left(q_{elec}\right)=E_{elec}\Psi_{elec}\left(q_{elec}\right) \tag{6}$$

$$\widehat{H}_{elec}=\sum_{k}^{elec}\hat{h}\left(r_k\right)+\sum_{kl,k>l}^{elec}r_{kl}^{-1}+V_{NN} \tag{7}$$

$$\hat{h}\left(r_k\right)=-\frac{1}{2}\nabla^2\left(r_k\right)-\sum_{\alpha}^{nuc}\frac{Z_\alpha}{R_{\alpha k}} \tag{8}$$

The electronic Hamiltonian consists of 1-electron operators $\hat{h}(r)$ that describe the kinetic energy of electron motion and the electrostatic interaction between electrons and nuclei. There is also a 2-electron operator that describes the pairwise electrostatic interactions between electrons. This study only focuses on the electronic Schrodinger equation, so the subscript "$elec$" will be dropped. The electronic Schrodinger equation will also be called Schrodinger equation for brevity. Two popular approaches to solve the Schrodinger equation are the variational and many-body perturbation theory methods. Our current research is restricted to stable closed-shell molecular systems. The solution of the Hartree-Fock (HF) equations,[18] which is a popular application of the variational method, is usually a good starting point to describe the ground state of these systems. For more accurate results, the second-order Moller-Plesset (MP2) perturbation theory,[19] which is a popular application of the many-body perturbation theory, can be carried out on top of the HF solution to improve the energy.

1.3.3 Hartree-Fock method

Following the variational method procedure, the HF method is formulated in three steps. First, a guessed wave function is built with a set of variational parameters and constraints. The energy expectation value of the trial wave function is evaluated, followed by energy minimization under the initial constraints. The last step is usually carried out using the Lagrange multiplier method. Since an N-electron wave function must be normalized and antisymmetric, in the HF method, it is best approximated by a single Slater determinant[20,21] $\Phi(x_1, x_2, \ldots, x_N)$, which is a determinant of N orthonormal 1-electron spin-orbital functions $\{\chi_p(x)\}$.[22]

$$\Phi\left(x_1, x_2, \cdots, x_N\right) = \frac{1}{\sqrt{N}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \cdots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \cdots & \chi_N(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \chi_1(x_N) & \chi_2(x_N) & \cdots & \chi_N(x_N) \end{vmatrix} \tag{9}$$

A spin-orbital $\chi_p(x)$, also called a molecular orbital (MO), is a product of a spatial MO $\varphi_p(r)$ and a spin function, which can be a spin up $[\alpha(\omega)]$ or spin down $[\beta(\omega)]$ function. Both sets of spatial orbitals and spin functions are usually chosen to be orthonormal.

$$\int dr \varphi_p^*(r) \varphi_q(r) = \delta_{pq} \tag{10}$$

$$\int dw a(w) a(w) = \int dw b(w) b(w) = 1 \tag{11}$$

$$\int dw a(w) b(w) = 0 \tag{12}$$

In equation (10), δ_{pq} is the delta Kronecker. In the restricted treatment, a set of K spatial MOs can be used to build a set of $2K$ spin-orbital by multiplying the spatial orbital with either spin up or a spin down function:[23]

$$\chi_{2p-1}(x) = \varphi_p(r)\alpha(\omega) \tag{13}$$

$$\chi_{2p}(x) = \varphi_p(r)\beta(\omega) \tag{14}$$

A restricted Slater determinant with all spatial MOs doubly occupied by electrons is usually a good trial wave function for stable closed-shell molecular systems, which are the target in our current studies. The energy expectation of the restricted closed-shell Slater determinant can be obtained from the Slater-Condon rules:[21,24]

$$E = \left\langle \Phi \middle| \widehat{H} \middle| \Phi \right\rangle = \iiint dq \Phi^*(q) \widehat{H} \Phi(q) = \sum_k^{occ} \left(2h_{kk} + J_{kk} - K_{kk} \right) \tag{15}$$

In Eq. (15), *occ* stands for the occupied spatial MOs. The energy expectation includes the 1-electron integrals h_{kk} that describe the kinetic energy of the electrons and their electrostatic interaction with the nuclei. The one-electron integrals are formulated in terms of the 1-electron operator defined in Eq. (8) and the spatial MOs as follows:

$$h_{pq} = \int dr \varphi_p^*(r) \hat{h}(r) \varphi_q(r) \tag{16}$$

The 2-electron integrals include the electron-electron classical Coulomb electrostatic interaction (J_{kk}), and non-classical exchange interaction (K_{kk}). These integrals can generally be defined in terms of Coulomb $\hat{J}(r)$ and exchange $\widehat{K}(r)$ operators and spatial MOs as follows

$$J_{pq} = \int dr_1 \varphi_p^*(r_1) \hat{J} \varphi_q(r_1) = 2\sum_k^{occ} \iint dr_1\, dr_2 \varphi_k^*(r_1) \varphi_k(r_1) \varphi_p^*(r_2) \varphi_q(r_2) = 2\sum_k^{occ} (pq \mid kk) \tag{17}$$

$$K_{pq} = \int dr_1 \varphi_p^*(r_1) \widehat{K} \varphi_q(r_1) = \sum_k^{occ} \iint dr_1\, dr_2 \varphi_k^*(r_1) \varphi_q(r_1) r_{12}^{-1} \varphi_p^*(r_2) \varphi_k(r_2) = \sum_k^{occ} (pk \mid kq) \tag{18}$$

Using the Lagrange multiplier method, the energy expectation can be minimized with the MO orthonormality constraint that leads to a pseudo-eigenvalue equation of the Fock operator

$$\hat{f}(r)\varphi_p(r) = \sum_q^{occ} \varepsilon_{pq}\varphi_q(r) \tag{19}$$

The Fock operator $\hat{f}(r)$ consists of the 1-electron operator $\hat{h}(r)$ defined in Eq. (8), the Coulomb $\hat{J}(r)$ and exchange $\hat{K}(r)$ operators in Eqs. (17) and (18).

$$\hat{f}(r) = \hat{h}(r) + \hat{J}(r) - \hat{K}(r) \tag{20}$$

Since the energy expectation is invariant to a unitary transformation, the MO basis can be rotated so that the Lagrange multiplier ε_{pq} becomes a diagonal matrix:

$$\hat{f}(r)\varphi_p(r) = \varepsilon_p\varphi_p(r) \tag{21}$$

The HF equation is usually solved numerically by expanding the MO $\varphi_p(r)$ as a linear combination of atomic orbital (LCAO) basis functions $\{\phi_\mu(r)\}$.[25]

$$j_p(r) = \sum_m^{AO} f_m(r) C_{mp} \tag{22}$$

The AO basis functions can be Slater-type functions[26] $f(r) \times e^{-\alpha r}$ or Gaussian-type functions[27] $f(r) \times e^{-\alpha r^2}$, which are pre-built and tabulated for most chemical elements. Therefore, the energy is minimized in the LCAO coefficient space $\{C_{\mu p}\}$. In the AO grid, the HF equation is converted into the Roothaan matrix equations[28]

$$FC = SCE \tag{23}$$

The Fock $F_{\mu\nu}$ and overlap $S_{\mu\nu}$ matrix elements are defined in Eqs. (24) and (28). The Fock matrix itself is a function of the LCAO coefficients; therefore, Eq. (23) has to be solved iteratively until the density matrix $D_{\mu\nu}$ in Eq. (27) is self-consistent.

$$F_{\mu\nu} = H_{\mu\nu} + \sum_{\mu\nu} (\mu\nu \mid \lambda\sigma) \left[D_{\mu\nu} - \frac{1}{2} D_{\lambda\sigma} \right] \tag{24}$$

$$H_{\mu\nu} = \int dr \phi_{\mu}^{*}(r) \hat{h}(r) \phi_{\nu}(r) \tag{25}$$

$$(mn \mid ls) = \iint dr_1 \, dr_2 f_m^*(r_1) f_n(r_1) r_{12}^{-1} f_l^*(r_2) f_s(r_2) \tag{26}$$

$$D_{\mu\nu} = 2 \sum_{k}^{occ} C_{\mu k} C_{\nu k} \tag{27}$$

$$S_{\mu\nu} = \int dr \phi_{\mu}^{*}(r) \phi_{\nu}(r) \tag{28}$$

Since the set of AO basis functions is not orthogonal (e.g., the overlap matrix S is not the identity matrix), the Roothaan equation is again a pseudo-eigenvalue equation. The first step to solve this equation is to rotate the AO basis to an orthonormal basis; i.e., finding a transformation matrix X so that the similarity transformation of the overlap matrix S makes an identity matrix.

$$X^{\dagger} S X = I \tag{29}$$

The Fock matrix and the LCAO vector matrix can subsequently be transformed as follows

$$C = X \tilde{C} \tag{30}$$

$$\tilde{F} = X^{\dagger} F X \tag{31}$$

The Roothaan equations become a matrix eigenvalue equation (Eq. (32)), which can be solved by diagonalizing the transformed Fock matrix $\tilde{F}$ for the eigenvector $\tilde{C}$ and eigenvalue matrix E. The LCAO coefficient matrix C can be obtained from Eq. (30).

$$\widetilde{F}\widetilde{C} = \widetilde{C}E \tag{32}$$

Finally, in terms of the AO basis and density matrix, the RHF energy expectation is given by

$$E = \sum_{\mu\nu}^{AO} H_{\mu\nu} D_{\mu\nu} + \frac{1}{2} \sum_{\mu\nu\lambda\sigma}^{AO} (\mu\nu \mid \lambda\sigma) \left(D_{\mu\nu} D_{\lambda\sigma} - \frac{1}{2} D_{\mu\lambda} D_{\nu\sigma} \right) \tag{33}$$

1.3.4 The second-order Moller-Plesset perturbation theory

Perturbation theory is a mathematical technique that solves an equation for a complex system in terms of a simple one. In quantum chemistry, perturbation theory starts with a simple reference (zeroth order) Hamiltonian $\hat{H}^{(0)}$, whose solution is known. When the difference ξ between the reference Hamiltonian $\hat{H}^{(0)}$ and the full Hamiltonian $\hat{H}$ is small, it is called a perturbation. The solution correction, which is difference between the reference solution and the exact solution, can be calculated in terms of the perturbation ξ and the reference solutions.

1.3.4.1 Rayleigh-Schrodinger Perturbation theory

In the general perturbation treatment for the Schrodinger equation,[29] the Hamiltonian $\hat{H}$ is split into the reference Hamiltonian operator $\hat{H}^{(0)}$ and the perturbation ξ

$$H = H^{(0)} + \lambda\xi \tag{34}$$

The parameter λ is set to 0 (off) or 1 (on). The eigenvalue equation of the reference Hamiltonian $\widehat{H}^{(0)}$ is solvable, thereby providing a complete set of orthonormal eigenfunctions $\{\Psi_i^{(0)}\}$ and the corresponding eigenvalues $\{E_i^{(0)}\}$.

$$H^{(0)}\Psi^{(0)} = E^{(0)}\Psi^{(0)} \tag{35}$$

Solutions of the exact Hamiltonian $\widehat{H}$ can be expanded in terms of the reference solutions. For instance, the *i*th state of the eigenfunction and eigenvalue are given by

$$\Psi_i = \Psi_i^{(0)} + \sum_m^{\infty} \lambda^m \Psi_i^{(m)} \tag{36}$$

$$E_i = E_i^{(0)} + \sum_m^{\infty} \lambda^m E_i^{(m)} \tag{37}$$

Plugging Eqs. (34),(36) and (37) into the amplitude Eq. (6), and equating terms with the same order of λ gives the corrected energy to all desired orders. For instance, the 1st – 3rd order correction energies are given by:

$$E_i^{(1)} = \left\langle \Psi_i^{(0)} \middle| \xi \middle| \Psi_i^{(0)} \right\rangle \tag{38}$$

$$E_i^{(2)} = \sum_{n\neq i}^{\infty} \frac{\left\| \left\langle \Psi_i^{(0)} \middle| \xi \middle| \Psi_n^{(0)} \right\rangle \right\|^2}{E_i^{(0)} - E_n^{(0)}} \tag{39}$$

$$E_i^{(3)} = \sum_{mn\neq i}^{\infty} \frac{\left\langle \Psi_i^{(0)} \middle| \xi \middle| \Psi_n^{(0)} \right\rangle \left\langle \Psi_n^{(0)} \middle| \xi \middle| \Psi_m^{(0)} \right\rangle \left\langle \Psi_m^{(0)} \middle| \xi \middle| \Psi_i^{(0)} \right\rangle}{\left(E_i^{(0)} - E_n^{(0)}\right)\left(E_i^{(0)} - E_m^{(0)}\right)} - E_i^{(i)} \sum_{n\neq i}^{\infty} \frac{\left\| \left\langle \Psi_i^{(0)} \middle| \xi \middle| \Psi_n^{(0)} \right\rangle \right\|^2}{\left(E_i^{(0)} - E_n^{(0)}\right)^2} \tag{40}$$

1.3.4.2 The second-order Moller-Plesset perturbation theory

In practice, perturbation theory can be applied to the Hartree-Fock mean field approximation, the Moller-Plesset approach to perturbation theory.[19] The reference Hamiltonian $\widehat{H}^{(0)}$ is defined as the *shifted* Fock operator

$$\widehat{H}^{(0)} = \widehat{F} + \left\langle \Phi^{(0)} \middle| \widehat{H} - \widehat{F} \middle| \Phi^{(0)} \right\rangle \tag{41}$$

The $\widehat{F}$ is the sum of the 1-electron Fock operators.

$$\widehat{F} = \sum_{k}^{elec} \widehat{f}\left(r_k\right) \tag{42}$$

The $\Phi^{(0)}$ is the normalized ground state Slater determinant wave function obtained from solving the Hartree-Fock equations (Eqs. (21)). Therefore, for the restricted closed-shell treatment, the expectation of $\widehat{F}$ is the sum of MO energy, and that of the exact Hamiltonian $\widehat{H}$ is the HF energy (see Eq. (15)):

$$\left\langle \Phi^{(0)} \middle| \widehat{F} \middle| \Phi^{(0)} \right\rangle = 2\sum_{k}^{occ} \varepsilon_k \tag{43}$$

$$\left\langle \Phi^{(0)} \middle| \widehat{H} \middle| \Phi^{(0)} \right\rangle = E^{HF} \tag{44}$$

With the reference Hamiltonian defined in Eq. (41), the perturbation operator is

$$\xi = \widehat{H} - \left(\widehat{F} + E^{HF} - 2\sum_{k}^{occ} \varepsilon_k \right) \tag{45}$$

Using Eq. (38), (43) and (45), the first-order energy correction is *zero*:

$$E^{(1)} = \left\langle \Phi^{(0)} \middle| \widehat{H} \middle| \Phi^{(0)} \right\rangle - \left\langle \Phi^{(0)} \middle| \widehat{F} \middle| \Phi^{(0)} \right\rangle - E^{HF} + 2\sum_{k}^{occ} \varepsilon_k = 0 \tag{46}$$

Using Eq. (39), the second-order energy correction is

$$E^{(2)} = \sum_{n\neq 0}^{\infty} \frac{\left\| \left\langle \Phi^{(0)} \middle| \xi \middle| \Phi_n^{(0)} \right\rangle \right\|^2}{E^{HF} - E_n^{(0)}} \tag{47}$$

The $\Phi_n^{(0)}$ are excitation determinants obtained by exchanging occupied MOs in $\Phi^{(0)}$ with virtual MOs. For instance, exchanging one occupied MO with one virtual MO introduces the singly excitation determinant; exchanging a pair of occupied MOs with a pair of virtual MOs gives doubly excitation MOs.

According to the Brillouin theorem, the cross term in the numerators of Eq. (47) due to the coupling with singly excitation determinant is zero. Using the Slater-Condon's rule, only terms with doubly excitation Slater determinant are non-zero. After some algebra, the second-order Moller-Plesset (MP2) energy correction can be formulated in terms of 4-2ERIs in the MO basis as follows:

$$E^{(2)} = \sum_{ij}^{occ} \sum_{ab}^{virt} (ia \mid jb) \left[2t_{ab}^{ij} - t_{ba}^{ij} \right] \tag{48}$$

$$t_{ab}^{ij} = \frac{(ia \mid jb)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{49}$$

1.3.5 Fragment molecular orbital method

As illustrated for the HF and MP2 methods, the main computational demands in *ab initio* electronic structure calculations are usually i) the evaluation and storage of two-electron integrals; ii) diagonalization of the Fock matrix; iii) transforming the integrals from the AO to the MO basis. The HF and MP2 computational costs scale as $O(N^{4-5})$, in which N measures the size

of the system, e.g., the number of atoms or the number of the AO basis functions. The computational cost (and memory demand) can be huge for macromolecular systems.

Many methods have been developed to treat macromolecules efficiently. One approach is to design small prototype systems to mimic active areas of macromolecules. However, such models usually omit long-range interactions and 3-D effects, such as stereochemistry, that can be critical in macrosystems. In a second approach, the surroundings around the active sites are included but treated at lower level of theory than the active site. For instance, the active sites might be treated with quantum mechanics (QM) while molecular mechanics (MM) is used for the environment. A third approach is to combine QM at high level of theory for active sites, and QM at lower level of theory for the surroundings.

Another class of methods designed to greatly expand the sizes of accessible systems is based on partitioning the system into fragments. Examples include the effective fragment potential (EFP) method,[30] the fragment molecular orbital (FMO) method,[1–5] and the effective fragment molecular orbital (EFMO) method.[11,12] The FMO and EFMO methods have been shown to scale linearly while preserving the accuracy of the underlying *ab initio* method. Since the gradient of the EFMO charge transfer term in EFMO is still being developed, research in this dissertation mainly focuses on the FMO method.

Since the Coulomb interaction is long-range while exchange is a short-range interaction, in the FMO method, the *ab initio* calculation for each fragment (monomer) is only embedded in the Coulomb electrostatic potential due to the electron density and nuclei of all other fragments. The other fragment-fragment interactions (e.g., exchange, charge transfer, induction) are subsequently accounted for by a many-body expansion method that requires calculations for

pairs (dimers) or triples (trimers) of fragments. The FMO energy with 2-body corrections, for instance, is given by:

$$E = \sum_{I} E^{I} + \sum_{IJ,J>I} \left(E^{IJ} - E^{I} - E^{J} \right) \tag{50}$$

The FMO calculation is started by solving the Hartree-Fock equation for monomers in the ESP of the other monomers. The monomer calculations are repeated until the densities of all monomers are converged. In the second stage of the calculations, dimers and/or trimers are embedded in *fixed* electrostatic fields of the converged monomers. The charge distribution of dimers or trimers is, therefore, generally different from the corresponding charge distribution of the monomers. Due to the monomer ESP, the Fock matrix of a fragment (X) consists of an internal fragment component $\tilde{F}^{X}$ and an ESP term $\bar{F}^{X}$.

$$F_{pq}^{X} = \widetilde{F}_{pq}^{X} + \overline{F}_{pq}^{X} \tag{51}$$

The internal Fock matrix element is formulated similarly to that of isolated molecules, including a 1-electron term $\widetilde{H}^{X}$ that describes the kinetic energy of the electrons and the electrostatic interaction between electrons and nuclei; the 2-electron term includes the Coulomb $\tilde{J}^{X}$ and exchange $\widetilde{K}^{X}$ integrals:

$$\widetilde{F}_{pq}^{X} = \widetilde{H}_{pq}^{X} + \widetilde{J}_{pq}^{X} - \widetilde{K}_{pq}^{X} \tag{52}$$

$$\widetilde{H}_{pq}^{X} = \left(p \left| -\frac{1}{2}\nabla^{2}\left(r_{k}\right) - \sum_{\alpha \in X}^{nucX} \frac{Z_{\alpha}}{R_{\alpha k}} \right| q \right) \tag{53}$$

$$\widetilde{J}_{pq}^{X} - \widetilde{K}_{pq}^{X} = \sum_{k \in X}^{occX} \left[2\left(pq \,|\, kk\right) - \left(pk \,|\, kq\right) \right] \tag{54}$$

In Eqs. (53) and (54), $nucX$ and $occX$ are nuclei and occupied MOs of the fragment X, respectively. In addition to terms defined in Eqs. (52)-(54), additional terms must be added to the internal Fock matrix elements when the fragmentation breaks covalent bonds. In the hybrid orbital projection (HOP) treatment,[3,31] the HOP contribution P_{pq}^{X} to the Fock matrix is defined in Eq. (55) through the HOP operator $\hat{P}^{X}$ in Eq. (56), in which θ_k is a hybrid orbital, and $B_k = 10^{6-8}$ is called the universal constant.[3,31]

$$P_{pq}^{X} = \left(p \middle| \hat{P}^{X} \middle| q \right) \tag{55}$$

$$\hat{P}^{X} = \sum_{k \in X} B_k \left| \theta_k \right\rangle \left\langle \theta_k \right| \tag{56}$$

The ESP Fock matrix element $\bar{F}_{pq}^{X}$ contains the Coulomb interactions between electrons in the fragment X with nuclei, $\bar{u}_{pq}^{X}$, and the electron density $\bar{J}_{pq}^{X}$ of all other fragments

$$\bar{F}_{pq}^{X} = \bar{u}_{pq}^{X} + \bar{J}_{pq}^{X} \tag{57}$$

$$\bar{u}_{pq}^{X} = \left(p \middle| - \sum_{K \neq X} \sum_{a \in K}^{nucK} \frac{Z_a}{R_{ak}} \middle| q \right) \tag{58}$$

$$\bar{J}_{pq}^{X} = 2 \sum_{K \neq X} \sum_{k \in K}^{occK} \left(pq \middle| kk \right) \tag{59}$$

In Eqs. (58) and (59), *nucK* and *occK* are the nuclei and occupied MOs of fragment $K \neq X$. Optimal MOs for fragment X are obtained by solving the Roothaan equations for the Fock matrix. In the FMO method, the Roothaan equations are solved iteratively until the densities of all monomers are self-consistent. Dimer and trimer contributions are calculated once the monomer density has converged. Finally, the fragment energy is given by:

$$E_X = \sum_{i \in X}^{occX} \left(\widetilde{H}_{ii}^{X} + \overline{F}_{ii}^{X} + F_{ii}^{X} \right) \tag{60}$$

Since the FMO method partitions a molecule into small fragments, it naturally facilitates multi-level parallelization. In the first level of parallelization, the computing resources can be distributed among fragments. The second level of parallelization is within each fragment calculation. The serial part in multi-level parallelization is thus eliminated, thereby enhancing the scalability of the FMO method. In GAMESS,[32] the multi-level parallelization is supported by the group distributed data interface (GDDI).[6]

1.3.6. Integral compressor

Memory storage and floating point operations for the transformation of 4-2ERIs from the AO to the MO basis in correlated methods (e.g., Eq. (48) in the MP2 method) is large. It is mandatory to reduce the dimension of the 4-2ERI matrix while retaining the accuracy of the calculations. In the early days of quantum chemistry, Ruedenberg,[33] Newton,[34] and Billingsley[35] approximated the two-center overlap density distribution as the sum of the squares of a one-center density distribution. For instance, for the set of four AOs $\phi_\mu, \phi_\nu, \phi_\lambda, \phi_\sigma$, the 2-center overlap charge distribution $\phi_\mu^* \phi_\nu$ and $\phi_\lambda^* \phi_\sigma$ was expanded in terms of the squares of AOs and proportional factors $Q_{\mu\nu}$ and $Q_{\lambda\sigma}$ as follows.

$$\phi_\mu^* \phi_\nu \approx Q_{\mu\nu} \left(\phi_\mu^* \phi_\mu + \phi_\nu^* \phi_\nu \right) \tag{61}$$

$$\phi_\lambda^* \phi_\sigma \approx Q_{\lambda\sigma} \left(\phi_\lambda^* \phi_\lambda + \phi_\sigma^* \phi_\sigma \right) \tag{62}$$

The 4-2ERI was, therefore, approximated as a linear combination of 2-2ERIs

$$(\mu\nu|\lambda\sigma) \approx Q_{\mu\nu} Q_{\lambda\sigma} \left[(\mu\mu|\lambda\lambda) + (\mu\mu|\sigma\sigma) + (\nu\nu|\lambda\lambda) + (\nu\nu|\sigma\sigma) \right] \tag{63}$$

Following these ideas, Whitten[36] established rigorous mathematical theorems on the error bounds for 4-2ERI evaluated with approximate densities. These theorems enabled tools to optimize the auxiliary basis that spans the space of approximate density. Using these theorems, Dunlap[37,38] and Almlöf[39] presented different schemes to minimize the residual density, which is the difference between the approximate and exact densities, introducing at least three approximations called SVS, S and V. The V-type approximation was found to be the most accurate one for 4-2ERI evaluation because it is based on the minimization of the Coulomb integral of the residual density.

The formulation of the V-type approximation starts with approximating the exact 2-center overlap density $\rho(r)$, which is the product of two AOs (Eq. (64)), by the approximate density $\tilde{\rho}(r)$ expanded in the auxiliary basis $\{\alpha_P(r)\}$ (Eq. (65)). The residual density $\Delta\rho(r)$ is then defined as the difference between the approximate $\tilde{\rho}(r)$ and the exact density $\rho(r)$.

$$r(r) = f_m^*(r) f_n(r) \tag{64}$$

$$\tilde{\rho}(r) = \sum_P^{aux} \alpha_P(r) C_P \tag{65}$$

$$\Delta\rho(r) = \rho(r) - \tilde{\rho}(r) \tag{66}$$

Minimizing the Coulomb integral of the residual density $(\Delta\rho|r_{12}^{-1}|\Delta\rho)$ in terms of the expansion coefficient C_P (eq. (65)), the 4-2ERI is optimally approximated by the product of 3-2ERIs $(\mu\nu|P)$ and 2-2ERIs V_{PQ}.

$$(\mu\nu\,|\,\lambda\sigma) \simeq \sum_{PQ}^{aux} (\mu\nu\,|\,P) V_{PQ}^{-1} (Q\,|\,\lambda\sigma) \tag{67}$$

$$(mn|P) = \iint dr_1\, dr_2 f_m^*(r_1) f_n(r_1) r_{12}^{-1} a_P(r_2) \tag{68}$$

$$V_{PQ} = \iint dr_1\, dr_2 a_P(r_1) r_{12}^{-1} a_Q(r_2) \tag{69}$$

The inverse of the matrix V can be decomposed and combined with the 3-2ERIs to form the 3-index matrix $\tilde{B}$ that can be used to form 4-2ERIs on-the-fly

$$V_{PQ}^{-1} = \sum_R^{aux} W_{PR} W_{RQ}^{\dagger} \tag{70}$$

$$\tilde{B}_{\mu\nu}^{P} = \sum_R^{aux} (\mu\nu|R) \Omega_{RP} \tag{71}$$

$$(\mu\nu|\lambda\sigma) \approx \sum_P^{aux} \tilde{B}_{\mu\nu}^{P} \tilde{B}_{\lambda\sigma}^{P,\dagger} \tag{72}$$

The matrix $\tilde{B}$ can also be transformed to the MO basis (Eq. (73)), which can be used to form 4-2ERIs in the MO basis (Eq. (74)). As only two AO indices need transforming, this introduces the main computational savings of the RI approximation (e.g., in correlation methods that need 4-2ERIs in the MO basis)

$$B_{pq}^{P} = \sum_{\mu\nu}^{ao} C_{\mu p} C_{\nu q} \tilde{B}_{\mu\nu}^{P} \tag{73}$$

$$(pq|rs) \approx \sum_P^{aux} B_{pq}^{P} B_{rs}^{P,\dagger} \tag{74}$$

1.3.7 MO response

1.3.7.1 Energy gradient

Besides the energy expectation, responses of a system (e.g., energy changes) to an internal or an external stimulation are relevant. For instance, the first-order energy changes with respect

to the nuclear displacement ζ determines the force acting on the nuclei. This quantity is called the nuclear gradient or energy gradient, which is essential for probing the potential energy surface; e.g., locating stationary points, determining the evolution of the system in molecular dynamic simulation.

In terms of the LCAO approximation, the energy is a function of 1- and 2-electron integrals in the AO basis, and of the LCAO coefficients. In the unperturbed state, e.g., $\zeta = 0$, denoted by the superscript (0), this can be represented as follows.

$$E^{(0)} = E^{(0)}\left(H_{mn}^{(0)};\left(mn \mid ls\right)^{(0)};C_{mp}^{(0)}\right) \tag{75}$$

In response to a perturbation (ζ), the unperturbed energy $E^{(0)}$ becomes the perturbed energy E, which is a function of perturbed LCAO coefficients and integrals of perturbed AOs.

$$E^{(0)} \mapsto E = E\left(H_{\mu\nu};\left(\mu\nu \mid \lambda\sigma\right);C_{\mu p}\right) \tag{76}$$

In this section, LCAO coefficients and matrix elements without the superscript (0) imply the perturbed state. The energy of the perturbed system can be expanded about the unperturbed state using the Taylor's series expansion

$$E = E^{(0)} + \sum_{i=1}^{\infty} \zeta^i E^{\zeta^i} \tag{77}$$

The superscript ζ^i stands for the i^{th}-order energy derivative with respect to ζ.

$$E^{\zeta^i} = \left(\frac{\partial^i}{\partial \zeta^i} E\right)_{\zeta=0} \tag{78}$$

The first derivative of the energy of the perturbed system with respect to the nuclear displacement E^{ζ} is the energy gradient. The energy gradient is apparently a function of the derivatives of LCAO coefficients C^{ζ} and integrals in the AO basis in the perturbed state:

$$E^{z} = E^{z}\left(H_{mn}^{z};\left(mn\,|\,s\right)^{z};C^{z}\right) \tag{79}$$

1.3.7.2 MO response

Since the AO basis functions are explicit functions of the nuclear coordinates, the derivatives of the 1- and 2-electron integrals in the AO basis with respect to the nuclear displacement ζ are known. However, differentiating LCAO coefficients is less straightforward. The perturbed LCAO coefficients are usually transformed into the unperturbed ones using a transformation matrix U.

$$C = C^{(0)}U \tag{80}$$

When the perturbation is off, the transformation matrix is simply an identity matrix I.

$$U_{\zeta=0} = I \tag{81}$$

The derivative of the perturbed LCAO coefficient becomes the derivative of the transformation matrix U, which is called the MO response matrix U^{ζ}

$$C^{\zeta} = C^{(0)}U^{\zeta} \tag{82}$$

1.3.7.3 The orthonormality equation of MO response

For the orthonormality choice of the MO basis, the overlap matrix in the MO basis is an identity matrix I [for both unperturbed and perturbed states]; that is

$$C^{(0),\dagger}SC^{(0)} = I \tag{83}$$

$$C^{\dagger}SC = I \tag{84}$$

Differentiating the perturbed MO overlap matrix gives

$$\left(\frac{\partial}{\partial z}C^{\dagger}SC\right)_{z=0} = 0 \tag{85}$$

Using Eqs. (82) and (81), the derivative of the perturbed MO overlap matrix becomes

$$U^{z,\dagger}C^{(0),\dagger}S^{(0)}C^{(0)} + C^{(0),\dagger}S^{z}C^{(0)} + C^{(0),\dagger}S^{(0)}C^{(0)}U^{z} = 0 \tag{86}$$

Using the MO orthonormality for unperturbed state (Eq. (83)), Eq. (86) is transformed into the well-known orthonormality condition for the MO response

$$U^{z,\dagger} + S^{(z)} + U^{z} = 0 \tag{87}$$

The overlap matrix with the nuclear displacement in the parentheses (ζ) stands for the derivative of the overlap matrix in the AO basis and then transformed back to the MO basis

$$S^{(\zeta)} = C^{(0),\dagger}S^{\zeta}C^{(0)} \tag{88}$$

1.3.7.4 Couple-perturbed Hartree-Fock equation

Equations for the MO response can be established from the Roothaan equations for the perturbed state

$$FC = SCE \tag{89}$$

Substituting perturbed LCAO coefficient matrix C by the unperturbed LCAO coefficient matrix $C^{(0)}$ (Eq. (80)) gives

$$FC^{(0)}U = SC^{(0)}UE \tag{90}$$

Multiplying the two sides of equation (91) by the transpose of the unperturbed LCAO coefficient matrix $C^{(0),\dagger}$ gives

$$\boldsymbol{\mathcal{F}}U = \boldsymbol{\mathcal{S}}UE \tag{91}$$

The bold letters stand for the transformed perturbed Fock and overlap matrices

$$\boldsymbol{\mathcal{F}} = C^{(0),\dagger}FC^{(0)} \tag{92}$$

$$\boldsymbol{\mathcal{S}} = C^{(0),\dagger}SC^{(0)} \tag{93}$$

Each perturbed quantity, say $\boldsymbol{\Lambda}$, in Eq. (91) can be expanded using the Taylor expansion similar to energy expansion in Eq. (77):

$$\Lambda = \Lambda^{(0)} + \sum_{i}^{\infty} \zeta^i \Lambda^{\zeta^i} \tag{94}$$

Collecting terms coupled with the first-order nuclear displacement introduces the couple-perturbed Hartree-Fock (CPHF) equation.

$$\boldsymbol{\mathcal{F}}^{\xi} + E^{(0)}U^{\xi} = \boldsymbol{\mathcal{S}}^{\xi}E^{(0)} + U^{\xi}E^{(0)} + E^{\xi} \tag{95}$$

$E^{(0)}$ is the diagonal matrix of the MO energy in the unperturbed state. E^{ζ} is the diagonal matrix of MO energy derivatives in the perturbed state. The specific form of the CPHF equations depends on the form of the Fock matrix, which will be discussed in the next chapters.

References

(1) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. Pair Interaction Molecular Orbital Method: An Approximate Computational Method for Molecular Interactions. *Chem. Phys. Lett.* **1999**, *312*, 319–324.

(2) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment Molecular Orbital Method: An Approximate Computational Method for Large Molecules. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(3) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment Molecular Orbital Method: Application to Polypeptides. *Chem. Phys. Lett.* **2000**, *318*, 614–618.

(4) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment Molecular Orbital Method: Use of Approximate Electrostatic Potential. *Chem. Phys. Lett.* **2002**, *351*, 475–480.

(5) Nagata, T.; Fedorov, D. G.; Kitaura, K. Mathematical Formulation of the Fragment Molecular Orbital Method BT - Linear-Scaling Techniques in Computational Chemistry and Physics: Methods and Applications; Zalesny, R., Papadopoulos, M. G., Mezey, P. G., Leszczynski, J., Eds.; Springer Netherlands: Dordrecht, 2011; pp 17–64.

(6) Fedorov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. A New Hierarchical Parallelization Scheme: Generalized Distributed Data Interface (GDDI), and an Application to the Fragment Molecular Orbital Method (FMO). *J. Comput. Chem.* **2004**, *25*, 872–880.

(7) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. The Distributed Data Interface in GAMESS. *Comput. Phys. Commun.* **2000**, *128*, 190–200.

(8) Pham, B. Q.; Gordon, M. S. Compressing the Four-Index Two-Electron Repulsion Integral Matrix Using the Resolution-of-the-Identity Approximation Combined with the Rank Factorization Approximation. *J. Chem. Theory Comput.* **2019**, *15*, 2254–2264.

(9) W., S. M.; K., B. K.; A., B. J.; T., E. S.; S., G. M.; H., J. J.; Shiro, K.; Nikita, M.; A., N. K.; Shujun, S.; et al. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **2004**, *14*, 1347–1363.

(10) de Lima Batista, A. P.; Zahariev, F.; Slowing, I. I.; Braga, A. A. C.; Ornellas, F. R.; Gordon, M. S. Silanol-Assisted Carbinolamine Formation in an Amine-Functionalized Mesoporous Silica Surface: Theoretical Investigation by Fragmentation Methods. *J. Phys. Chem. B* **2016**, *120*, 1660–1669.

(11) Steinmann, C.; Fedorov, D. G.; Jensen, J. H. Effective Fragment Molecular Orbital Method: A Merger of the Effective Fragment Potential and Fragment Molecular Orbital Methods. *J. Phys. Chem. A* **2010**, *114*, 8705–8712.

(12) Bertoni, C.; Gordon, M. S. Analytic Gradients for the Effective Fragment Molecular Orbital Method. *J. Chem. Theory Comput.* **2016**, *12*, 4743–4767.

(13) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.* **1926**, *28*, 1049–1070.

(14) Heisenberg, W. *The Physical Principles of the Quantum Theory*; Ill., The University of Chicago Press, 1930.

(15) Born, M. Statistical Interpretation of Quantum Mechanics. *Science (80-.).* **1955**, *122*, 675 LP – 679.

(16) Maurice, D. P. A.; Howard, F. R. Quantum Mechanics of Many-Electron Systems. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character* **1929**, *123*, 714–733.

(17) BORN, M.; OPPENHEIMER, R. ON THE QUANTUM THEORY OF MOLECULES. In *Quantum Chemistry*; World Scientific Series in 20th Century Chemistry; WORLD SCIENTIFIC, 2000; Vol. Volume 8, pp 1–24.

(18) Rayner, H. D.; W., H. Self-Consistent Field, with Exchange, for Beryllium. *Proc. R. Soc. London. Ser. A - Math. Phys. Sci.* **1935**, *150*, 9–33.

(19) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

(20) Maurice, D. P. A.; Howard, F. R. On the Theory of Quantum Mechanics. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character* **1926**, *112*, 661–677.

(21) Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **1929**, *34*, 1293–1322.

(22) Mulliken, R. S. Electronic Structures of Polyatomic Molecules and Valence. II. General Considerations. *Phys. Rev.* **1932**, *41*, 49–71.

(23) Szabó, A.; Ostlund, N. S. Modern Quantum Chemistry : Introduction to Advanced Electronic Structure Theory. 1996.

(24) Condon, E. U. The Theory of Complex Spectra. *Phys. Rev.* **1930**, *36*, 1121–1133.

(25) Mulliken, R. S. Spectroscopy, Molecular Orbitals, and Chemical Bonding. *Science (80-.).* **1967**, *157*, 13 LP – 24.

(26) Slater, J. C. Atomic Shielding Constants. *Phys. Rev.* **1930**, *36*, 57–64.

(27) F., B. S.; Charles, E. A. Electronic Wave Functions - I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. London. Ser. A. Math. Phys. Sci.* **1950**, *200*, 542–554.

(28) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

(29) McWeeny, R. *Methods of Molecular Quantum Mechanics*; Academic Press: London, 1992.

(30) 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.

(31) Nagata, T.; Fedorov, D. G.; Kitaura, K. Importance of the Hybrid Orbital Operator Derivative Term for the Energy Gradient in the Fragment Molecular Orbital Method. *Chem. Phys. Lett.* **2010**, *492*, 302–308.

(32) The General Atomic and Molecular Electronic Structure System (GAMESS 2018-R3). https://www.msg.chem.iastate.edu/gamess/.

(33) Rüdenberg, K. On the Three- and Four-Center Integrals in Molecular Quantum Mechanics. *J. Chem. Phys.* **1951**, *19*, 1433–1434.

(34) Newton, M. D. Self-Consistent Molecular-Orbital Methods. II. Projection of Diatomic Differential Overlap (PDDO). *J. Chem. Phys.* **1969**, *51*, 3917–3926.

(35) Billingsley, F. P.; Bloor, J. E. Limited Expansion of Diatomic Overlap (LEDO): A Near-Accurate Approximate Ab Initio LCAO MO Method. I. Theory and Preliminary Investigations. *J. Chem. Phys.* **1971**, *55*, 5178–5190.

(36) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(37) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. On Some Approximations in Applications of Xα Theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(38) Mintmire, J. W.; Dunlap, B. I. Fitting the Coulomb Potential Variationally in Linear-Combination-of-Atomic-Orbitals Density-Functional Calculations. *Phys. Rev. A* **1982**, *25*, 88–95.

(39) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral Approximations for LCAO-SCF Calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

CHAPTER 2. COMPRESSING THE FOUR-INDEX TWO-ELECTRON REPULSION INTEGRAL MATRIX USING THE RESOLUTION-OF-THE-IDENTITY APPROXIMATION COMBINED WITH THE RANK FACTORIZATION APPROXIMATION

A paper published in The Journal of Chemical Theory and Computation

Buu Q. Pham and Mark S. Gordon

Abstract

The four-index two-electron repulsion integral (4-2ERI) matrix is compressed using the resolution-of-the-identity (RI) approximation combined with the rank factorization approximation (RFA). The 4-2ERI is first approximated by the RI product. Then, the singular value decomposition (SVD) approximation is used to eliminate low-weighted singular vectors. The SVD RI approximation maintains the canonical form of the RI approximation and introduces a tunable compression factor. The characteristics of the SVD RI approximation along with the stochastic RI and natural auxiliary function approximation were numerically examined by applying these methods to the closed-shell second-order Moller-Plesset perturbation theory (MP2). The results show that while the SVD RI approximation yields large errors for absolute properties (e.g., the correlation energy), it provides accurate relative properties (potential energy surface, binding energy) of the applied *ab initio* method (e.g., RHF, MP2).

2.1 Introduction

The 4-index 2-electron repulsion integral (4-2ERI) is the kernel and the bottleneck of most *ab initio* electronic structure methods. Given a molecule specified by an atomic orbital (AO) basis $\{\phi_\mu(r)\}$, the 4-2ERI can be formulated as:

$$(\mu\nu\,|\,\lambda\sigma) = \iint dr_1 dr_2 \phi_\mu^*(r_1)\phi_\nu(r_1) r_{12}^{-1} \phi_\lambda^*(r_2)\phi_\sigma(r_2) \tag{1}$$

For an AO basis of *N* functions, full storage of 4-2ERIs would need an array of ~$O(N^4)$, which usually goes beyond the memory capacity of most single compute nodes for molecules of moderate size. Therefore, the 4-2ERI tensor is stored (if needed) in distributed memory arrays, or worse, on hard drives resulting in significant communication overhead in large-scale calculations. For correlation methods, the 4-2ERI tensor also must be partly to fully transformed from the AO to the molecular orbital (MO) basis, which requires expensive matrix multiplication operations.

Many techniques have been developed to reduce the computational cost of 4-2ERI evaluations. These include the resolution-of-the-identity (RI) approximation also called the density fitting (DF) approximation,[1–3] tensor hyper-contraction DF,[4–7] low-rank factorization,[8] and the Cholesky decomposition (CD).[9–12] The latter is based on the lower-bound of a positive definite operator,[13] which was first used to decompose the 4-2ERI tensor by Beebe and Linderberg in 1977,[14] and with the analytic gradient by Aquilante, Lindh, and Pedersen in 2008.[15] As the large 4-2ERI tensor is needed, and the resulting Cholesky vectors are also large, the CD is usually used in combination with other techniques like the RI,[16,17] the singular value decomposition,[18] and the Cuthill-McKee sparse matrix reordering.[19] Common among these methods is to factorize the 4-

2ERI tensor into a product of low dimensional tensors. Note that such 4-2ERI tensor factorizations are feasible because a matrix *A* of *M* rows, *N* columns and rank *r* can always be (accurately) decomposed into the following product.

$$A_{M'N} = W_{M'r} H_{r'N} \tag{2}$$

A benefit is gained when the rank *r* is significantly smaller than the dimension of the matrix *A*, e.g., $r \ll min(M, N)$. For large calculations, the rank *r* is made small (i.e., $r \to \tilde{r}$; $\tilde{r} \leq r$) to reduce the computational cost. The rank reduction introduces an approximation to Eq. (2) called the rank factorization approximation (RFA):

$$A_{M\times N} \simeq W_{M\times \tilde{r}} H_{\tilde{r}\times N} \tag{3}$$

The singular value decomposition (SVD) technique is the standard tool for the RFA, which can be carried out in two steps. First, the matrix *A* is decomposed (Eq. (4)) into the product of the left (U) and right (L) singular vector matrices and the singular value matrix (S).

$$A_{M'N} = U_{M'r} S_{r'r} L^{\dagger}_{r'N} \tag{4}$$

Here, *S* is a diagonal matrix of the singular values, which are sorted in descending order. The dagger $(\dagger)$ indicates the matrix transpose operation. The number of non-zero singular values in *S* is the rank *r* of the matrix *A*. The left and right singular matrices are orthonormal:

$$U^{\dagger}_{r'M} U_{M'r} = L^{\dagger}_{r'N} L_{N'r} = I_{r'r} \tag{5}$$

In Eq. (5), I is the identity matrix. The *exact* SVD decomposition of the matrix *A* is illustrated in FIGURE 2.1a. The singular value matrix *S* can be absorbed into the left or right singular vector

matrix to produce the rank factorization form shown in Eq. (3). Since the singular vectors that correspond to the larger singular values in the matrix *S* have higher weights in the SVD product than those that correspond to the smaller singular values in *S*, the best RFA can be achieved by *systematically* eliminating singular vectors whose singular values are smaller than an SVD threshold θ_{SVD} as shown in Eq. (6) and visualized in FIGURE 2.1b.

$$A_{M\times N} \simeq U_{M\times \tilde{r}} S_{\tilde{r}\times \tilde{r}} L^{\dagger}_{\tilde{r}\times N}; \left(S_{ii}\right)_{i\leq \tilde{r}} \geq \theta_{SVD} \tag{6}$$

The best RFA means that the Frobenius norm of the difference between the approximate and exact matrix $(A - USL^{\dagger})$ is minimized. The Frobenius norm[20] of a matrix Ξ is the square root of the trace of the product of that matrix with its transpose $\sqrt{Tr(\Xi\Xi^{\dagger})}$.

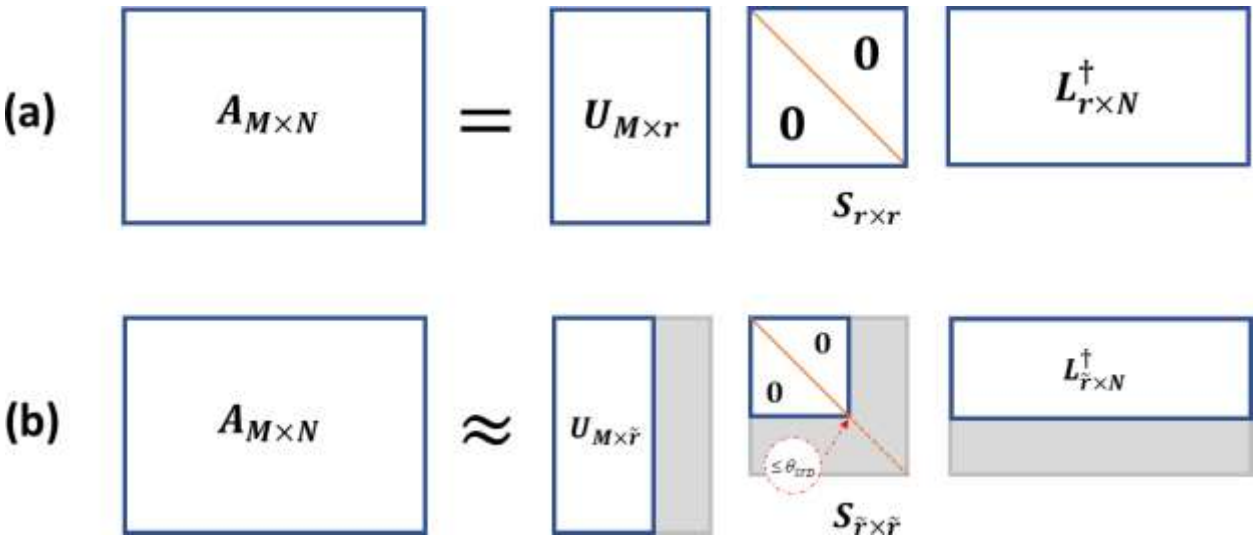


FIGURE 2.1 *(a) The exact and (b) approximate SVD factorization of the matrix $A_{M\times N}$. U, L and S are the left and right eigenvector matrices and the singular value matrix, respectively.*

Besides the direct RFA, a popular scheme to factorize the 4-2ERI tensor is to use the RI approximation,[1–3] in which, the factorization is accomplished through a molecular property; e.g., by fitting the overlap density distribution of the AO basis functions to an auxiliary basis. In terms of the RI approximation, the 4-2ERI matrix $G_{N^2\times N^2}$ is approximated by

$$G_{N^2\times N^2} \simeq B_{N^2\times k} B^{\dagger}_{k\times N^2} \tag{7}$$

In Eq. (7), the *four*-dimensional $N \times N \times N \times N$ 4-2ERI tensor, whose elements are defined in Eq. (1), has been reorganized into a *two*-dimensional matrix $G_{N^2\times N^2}$. In Eq. (7) *k* is the number of linearly independent auxiliary basis functions. The RI approximation can significantly reduce the size of the 4-2ERI matrix, e.g., from ~ $O(N^4)$ to ~ $O(N^2 k)$, usually with a small error. A drawback of the RI approximation is that it is a "fixed accuracy" method; i.e., the accuracy (and the computational cost) of the RI approximation can only be tuned by changing the auxiliary basis set. A tunable parameter that governs the accuracy and the computational cost while maintaining the simple form of the RI approximation has been introduced by Takeshita *et* al.[21] and Kallay.[22]

The core of the Takeshita *et* al.[21] stochastic RI approximation (STO RI) is a stochastic rectangular matrix $\Theta_{\tilde{k}\times k}$, whose elements are random numbers equal to +1 or -1 scaled by a factor of $\left(\tilde{k}\right)^{-1/2}$.

$$\Theta_{\tilde{k}\times k} = \left(\tilde{k}\right)^{-1/2} \begin{pmatrix} \pm 1 & \pm 1 & \cdots & \pm 1 \\ \pm 1 & \pm 1 & \cdots & \pm 1 \\ \cdots & \cdots & \cdots & \cdots \\ \pm 1 & \pm 1 & \cdots & \pm 1 \end{pmatrix} \tag{8}$$

The number of the stochastic matrix columns *k* is the number of linearly independent auxiliary basis functions. Each *k*-element row of the stochastic matrix is called a stochastic orbital (i.e., $\tilde{k}$ is the number of stochastic orbitals). Statistically, the stochastic matrix is expected to be an approximately orthonormal matrix; i.e., the product of the stochastic matrix with its transpose approximates the identity matrix:[21]

$$\Theta^{\dagger}_{k\times\tilde{k}}\Theta_{\tilde{k}\times k} \simeq I_{k\times k} \tag{9}$$

By inserting this approximation into the RI product (Eq. (7)), and combining the matrix *B* with the stochastic matrix Θ gives

$$G_{N^2\times N^2} \simeq \left(B\Theta^{\dagger}\right)_{N^2\times\tilde{k}}\left(B\Theta^{\dagger}\right)^{\dagger}_{\tilde{k}\times N^2} \tag{10}$$

The computational cost and the *accuracy* of the stochastic RI approximation can be controlled by the compression factor $k/\tilde{k}$, which can be tuned by varying the number of stochastic orbitals. While the stochastic RI approximation introduces a low *absolute* error *per* electron for large molecules,[21] its non-deterministic nature cannot even *qualitatively* reproduce relative properties (e.g., the potential energy surface (PES), binding energy), as the error tends to accumulate instead of cancelling out. One *non*-deterministic component of the stochastic RI approximation is due to the random elements of the stochastic matrix. A second *non*-deterministic factor is identical to the natural auxiliary function approach described in the following paragraphs.

In the natural auxiliary function (NAF) method,[22] an orthonormal matrix is generated by diagonalizing the product of the matrix *B* in Eq. (7) with its transpose as shown in Eq. (11). This can be done because the product of any matrix (*B* in this context) with its transpose is symmetric.

Diagonalizing the symmetric product $BB^{\dagger}$ (or $B^{\dagger}B$) introduces the left (or right) singular vector matrix and the square of the singular value matrix (of the matrix *B*, in this case). In the NAF approach, the matrix product $B^{\dagger}B$ is diagonalized (Eq. (11)) to form the orthonormal (Eq. (12)) right singular matrix *L* of the matrix *B*.

$$B^{\dagger}_{k' N^2} B_{N^{2'} k} = L^{\dagger}_{k' k} \mathsf{L}_{k' k} L_{k' k} \tag{11}$$

$$L^{\dagger}_{k' k} L_{k' k} = I_{k' k} \tag{12}$$

The matrix *L* is then row-truncated $L_{k\times k} \rightarrow L_{\tilde{k}\times k}$ to form an approximate orthonormal matrix as shown in Eq. (13). Inserting Eq. (13) into the RI product (Eq. (7)) introduces the final form of the NAF approximation (Eq. (14)).[22]

$$L^{\dagger}_{k\times\tilde{k}} L_{\tilde{k}\times k} \simeq I_{k\times k} \tag{13}$$

$$G_{N^2\times N^2} \simeq \left(BL^{\dagger}\right)_{N^2\times\tilde{k}} \left(BL^{\dagger}\right)^{\dagger}_{\tilde{k}\times N^2} \tag{14}$$

Examining the set of Eqs. (11)-(14), the NAF approach might appear to be a deterministic procedure. The row-truncated right singular matrix *L* in Eq. (13) was interpreted as a transformation matrix of the conventional auxiliary basis to a *natural auxiliary basis* (NAF). However, in terms of the RFA, the compression of the matrix *B* using the truncated orthonormal matrix *L*; i.e., $B_{N^2\times k} \rightarrow (BL^{\dagger})_{N^2\times\tilde{k}}$, has *arbitrarily* removed a set of $(k-\tilde{k})$ vectors from the matrix *B* with no knowledge of their weights. This implies that the NAF method is *non*-deterministic, which can destroy the correlation between the compression factor $k/\tilde{k}$ and the accuracy of the approximation (Eq. (14)).

The crucial step in the stochastic and the NAF RI approximations is to generate an approximate orthonormal matrix $\Gamma_{k\times\tilde{k}}$ with $\tilde{k} \leq k$. This can also be done by simply generating a random set of $\tilde{k}$ vectors, which are then orthonormalized using a Gram-Schmidt routine.[23] This set of orthonormal vectors can be used to form a matrix that is equivalent to the stochastic matrix (in the stochastic RI approximation) or the truncated right singular vector matrix (in the NAF approximation). In comparison with the NAF approach, this procedure avoids forming and diagonalizing the product $B^{\dagger}B$ (Eq. (11)). However, it is also a *non*-deterministic approach.

In the present paper, by using the SVD approximation (Eq. (6)), a proper RFA will be applied on top of the RI approximation to introduce a compression factor that can truly control the accuracy (and the computational cost) of the 4-2ERI evaluation. The numerical behavior of the new approach, called the SVD RI approximation, along with the double sampling stochastic RI approximation[21] and NAF method[22] will be examined by applying these methods to the closed-shell second-order Moller-Plesset perturbation theory (MP2) for water clusters, the T-shaped benzene dimer, and the S22 non-covalent complex test set.[24] The scaling of the SVD RI approximation with fully optimized implementation and parallelization will be presented in a later paper.

2.2 SVD-RI approximation

Using the Coulomb metric (also called the *V*-type) RI approximation,[3] the 4-2ERI shown in Eq. (1) can be approximated by the product of 3-2ERIs and 2-2ERIs:

$$\left(\mu\nu\,|\,\lambda\sigma\right) \simeq \sum_{PQ}^{X}\left(\mu\nu\,|\,P\right)V_{PQ}^{-1}\left(Q\,|\,\lambda\sigma\right) \tag{15}$$

Here, *X* is the number of auxiliary basis functions. The 3-index integrals are the ERIs of *two* AO basis functions μ,ν or λ,σ and *one* auxiliary basis function $\{\alpha_P\}$, while the 2-index integrals are the ERIs of *two* auxiliary basis functions:

$$(\mu\nu\,|\,P) = \iint dr_1 dr_2 \phi_\mu^*(r_1)\phi_\nu(r_1) r_{12}^{-1} \alpha_P(r_2) \tag{16}$$

$$V_{PQ} = \iint dr_1 dr_2 \alpha_P^*(r_1) r_{12}^{-1} \alpha_Q(r_2) \tag{17}$$

For the AO and auxiliary bases containing *N* and *X* functions, respectively, the 4-2ERIs, 3-2ERIs and 2-2ERIs can be put into the matrices $G_{N^2\times N^2}$, $K_{N^2\times X}$ and $V_{X\times X}$, respectively. Eq. (15) can subsequently be rewritten in matrix form as follows

$$G_{N^2\times N^2} \simeq K_{N^2\times X} V^{-1}_{X\times X} K^{\dagger}_{X\times N^2} \tag{18}$$

The inverse of the matrix *V* can be decomposed into the product of a matrix W and its transpose $\Omega^{\dagger}$ using the *LU* decomposition (e.g., the eigenvalue or the Cholesky decomposition method)

$$V^{-1}_{X\times X} = \Omega_{X\times k}\Omega^{\dagger}_{k\times X} \tag{19}$$

In Eq. (19), *k* is the rank of the inverse of the matrix *V*, which is also the number of linearly independent auxiliary basis functions. Combining the *V*-decomposed matrix $\Omega_{X\times k}$ with the matrix of 3-2ERIs $K_{N^2\times X}$ yields the fundamental matrix *B* of the RI approximation

$$B_{N^2\times k} = K_{N^2\times X}\Omega_{X\times k} \tag{20}$$

Using the matrix *B* and its transpose, the previously approximated 4-2ERI in Eq. (7) is reproduced:

$$G_{N^2\times N^2} \simeq B_{N^2\times k} B^{\dagger}_{k\times N^2} \tag{21}$$

The matrix *B* (or equivalently, its transpose) can be factored using the SVD:

$$B_{N^{2'}\,k} = U_{N^{2'}\,k'} S_{k'\,k'} L^{\dagger}_{k'\,k} \tag{22}$$

In Eq. (22), k' is the rank of the matrix *B*, which is the number of *non*-zero singular values. The two matrices *U* and L are the orthonormal left and right singular vector matrices

$$U^{\dagger}_{k'\times N^2} U_{N^2\times k'} = L^{\dagger}_{k'\times k} L_{k\times k'} = I_{k'\times k'} \tag{23}$$

The RFA is applied to the factorization in Eq. (22) by introducing a singular value threshold θ_{SVD} that leads to the truncation of columns and/or rows of the singular value and singular vector matrices as follows

$$S_{k'\times k'} \rightarrow S_{\tilde{k}\times\tilde{k}} \tag{24}$$

$$U_{N^2\times k'} \rightarrow U_{N^2\times\tilde{k}} \tag{25}$$

$$L^{\dagger}_{k'\times k} \rightarrow L^{\dagger}_{\tilde{k}\times k} \tag{26}$$

In Eqs. (24)-(26), $\tilde{k}\left(\leq k'\right)$ is the truncated dimension of the matrices *S*, *U* and *L* obtained by applying the threshold θ_{SVD} to the singular value matrix. The SVD approximation of the matrix *B* is given by:

$$B_{N^2\times k} \simeq U_{N^2\times\tilde{k}} S_{\tilde{k}\times\tilde{k}} L^{\dagger}_{\tilde{k}\times k} \tag{27}$$

Similarly, the SVD approximation of the transpose of the matrix *B* is:

$$B^{\dagger}_{k\times N^2} \simeq L_{k\times\tilde{k}} S^{\dagger}_{\tilde{k}\times\tilde{k}} U^{\dagger}_{\tilde{k}\times N^2} \tag{28}$$

Plugging Eqs. (27) and (28) into Eq. (21), the approximate 4-2ERI matrix becomes

$$G_{N^2\times N^2} \simeq \left(U_{N^2\times\tilde{k}} S_{\tilde{k}\times\tilde{k}}\right)\left(L^{\dagger}_{\tilde{k}\times k} L_{k\times\tilde{k}}\right)\left(S^{\dagger}_{\tilde{k}\times\tilde{k}} U^{\dagger}_{\tilde{k}\times N^2}\right) \tag{29}$$

Since the right singular matrix L is an orthonormal matrix, its column-truncated matrix $L_{k\times\tilde{k}}$ is also an *exact* orthonormal matrix; that is:

$$L^{\dagger}_{\tilde{k}\times k} L_{k\times\tilde{k}} = I_{\tilde{k}\times\tilde{k}} \tag{30}$$

Absorbing the singular values into the truncated left singular vector matrix U yields the fundamental matrix $\tilde{B}$ of the SVD RI approximation:

$$\widetilde{B}_{N^2\times\tilde{k}} = U_{N^2\times\tilde{k}} S_{\tilde{k}\times\tilde{k}} \tag{31}$$

Using the matrix $\tilde{B}$ and Eq. (30), Eq. (29) becomes

$$G_{N^2\times N^2} \simeq \widetilde{B}_{N^2\times\tilde{k}} \widetilde{B}^{\dagger}_{\tilde{k}\times N^2} \tag{32}$$

The matrix $\tilde{B}$ can also be transformed into the MO basis, which is used to form the 4-2ERIs in the MO basis if needed. The accuracy (and computational cost) of the SVD RI approximation (Eq. (32)) compared with the conventional RI approximation (Eq. (21)) is determined by the SVD threshold θ_{SVD}, or equivalently by the compression factor $k/\tilde{k}$, in which *k* is the number of linearly independent auxiliary basis functions, and $\tilde{k}$ is the SVD truncated dimension (cf. Eqs. (24) -(26)). In the next sections, the characteristics of the SVD RI approximation along with the

stochastic RI and the NAF approximations are numerically examined using either the full closed shell MP2 method or the conventional RI-MP2 method as a reference.

2.3 SVD RI-MP2 correlation energy

To facilitate the comparison, the SVD, NAF, stochastic, and the conventional RI approximations are briefly summarized. First, the 4-2ERI is approximated by the conventional RI approximation as follows (see Eqs. (15)-(21))

$$G_{N^2\times N^2} \simeq B_{N^2\times k} B^{\dagger}_{k\times N^2} \tag{33}$$

In Eq. (33), *k* is the number of linearly independent auxiliary basis functions. In terms of the SVD, NAF and stochastic RI approximations, the 4-2ERI matrix is approximated by a similar canonical formula:

$$G_{N^2\times N^2} \simeq \tilde{B}^{W}_{N^2\times \tilde{k}} \tilde{B}^{W,\dagger}_{\tilde{k}\times N^2} \tag{34}$$

In Eq. (34), *W* stands for the SVD, NAF or STO RI approximation. Depending on the approximation *W*, $\tilde{k}$ can be the number of stochastic orbitals (for the STO RI approximation, Eq. (8)), the number of rows of the *row*-truncated *right* singular vector matrix (NAF RI approximation, Eq. (13)), or the number of columns in the *column*-truncated *left* singular vector matrix (SVD RI approximation, Eq. (25)). The ratio $k/\tilde{k}$ is the compression factor. For the STO approximation, the fundamental matrix $\tilde{B}^{STO}_{N^2\times\tilde{k}}$ is given by

$$\tilde{B}^{STO}_{N^2\times\tilde{k}} = \left(B\Theta^{\dagger}\right)_{N^2\times\tilde{k}} \tag{35}$$

The stochastic matrix $\Theta_{\tilde{k}\times k}$ is defined in Eq. (8). For the NAF, the matrix $\tilde{B}^{NAF}_{N^2\times\tilde{k}}$ is defined through the row-truncated right singular vector matrix of the matrix *B* (see Eqs. (11)-(14)):

$$\tilde{B}^{NAF}_{N^2\times\tilde{k}} = \left(BL^{\dagger}\right)_{N^2\times\tilde{k}} \tag{36}$$

For the SVD approximation, the matrix $\tilde{B}^{SVD}_{N^2\times\tilde{k}}$ is defined through the column-truncated left singular vector and singular value matrices of the matrix *B* (see Eqs. (22)-(32)):

$$\tilde{B}^{SVD}_{N^2\times\tilde{k}} = U_{N^2\times\tilde{k}}S_{\tilde{k}\times\tilde{k}} \tag{37}$$

The MP2 dynamic correlation energy $\left(E^{(2)}\right)$ obtained by correlating the electrons in the active occupied MOs (*act*) using the virtual MOs (*virt*) is formulated as follows:

$$E^{(2)} = \sum_{ij}^{act}\sum_{ab}^{virt} Q^{ij}_{ab}\left(2A^{ij}_{ab} - A^{ij}_{ba}\right) \tag{38}$$

$$Q^{ij}_{ab} = \sum_{\mu}^{N} C_{\mu i}\sum_{\nu}^{N} C_{\nu a}\sum_{\lambda}^{N} C_{\lambda j}\sum_{\sigma}^{N}(\mu\nu\,|\,\lambda\sigma)C_{\sigma b} \tag{39}$$

$$A^{ij}_{ab} = \frac{Q^{ij}_{ab}}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{40}$$

Q^{ij}_{ab} is the 4-2ERI matrix in the MO basis; $C_{\mu i}$ is the MO coefficients; ε_p is the energy of the p^{th} MO. The indices *i*, *j* and *a*, *b* stand for active occupied and virtual MOs, respectively.

A pilot *three*-step implementation of the RI-MP2, SVD RI-MP2, double sampling STO RI-MP2 and NAF RI-MP2 approximations is shown in SCHEME 1. In *Step 1*, the matrix of 2-2ERIs $(V_{X\times X})$ is formed, inverted and decomposed into the matrix $\Omega_{X\times k}$ **(1.1-1.3)**. The 3-2ERIs are calculated and

stored in the matrix $K_{N^2\times X}$, which is then combined with $\Omega_{X\times k}$ to form the matrix $B_{N^2\times k}$ **(1.4-1.5)**.

In *Step 2*, the SVD approximation is applied to the matrix *B* (e.g., by using the DGESVD routine in the Intel MKL library **(2.1)**). For the SVD RI approximation, only the truncated left singular vectors are requested in the output and overwritten to the available matrix *B*, which is then scaled by the singular values to form $\tilde{B}^{SVD}$ **(2.2)**. For the NAF approximation, only the truncated right singular vectors are determined, and combined with the matrix *B* to form $\tilde{B}^{NAF}$ **(2.3)**. For the STO RI approximation, the stochastic matrix is generated using a random number generator and combined with the matrix *B* to form $\tilde{B}^{STO}$ **(2.4)**. The matrix *B* is then transformed into the MO basis (i.e., the active occupied – virtual block) using the MO coefficient matrix *C* via **(2.5)** and/or **(2.6)**.

In *Step 3*, for each pair of active occupied MOs, the matrix *B* (for the regular RI approximation **(3.1)**) or $\tilde{B}^{W}$ (for the SVD, NAF, and STO RI approximations **(3.2)**) is used to form the 4-2ERI in the MO basis (e.g., Q_{VV}^{ij} in Eq. (39)), followed by the formation of A_{VV}^{ij} **(3.3)** and the MP2 correlation energy **(3.4)**. Note that for the STO RI-MP2 approach, *two* random sets of the matrix $\tilde{B}^{STO}$ are formed. The first $\tilde{B}^{STO}$ is used to form Q_{ab}^{ij}, the second for A_{ab}^{ij} as shown in Eq. (38).

The SVD approximation was briefly mentioned by Kallay[22] who expressed regarding memory requirements. While memory concerns are not the main focus of the present work, the memory issue can be addressed based on the following factors: i) the computer memory has been significantly improved during the last decade; ii) in parallel computing, it is not necessary to hold a matrix in one compute node, rather it can be split and distributed to many nodes; and iii) the

recent popular hybrid distributed/shared memory (e.g., MPI/OpenMP) parallel model has significantly reduced the replicated process memory while supporting very large on-node shared memory data structures. By using the MPI/OpenMP model in the newly developed RI-MP2 gradient implementation, one could indeed load the entire or a large part of matrix B (see Eq. 21) into the process memory for a reasonably large system of 200 atoms on what are now commonly used chips (e.g., Haswell, KNL). The source code of the implementation can be found in the recently released version of GAMESS.[25]

SCHEME 2.1 *SVD RI MP2 correlation energy implementation*

Step 1. Form B for the RI approximation

1.1 $V_{X\times X} \leftarrow (P|Q)$

1.2 $V^{-1}_{X\times X} \leftarrow V_{X\times X}$

1.3 $V^{-1}_{X\times X} = \Omega_{X\times k}\Omega^{\dagger}_{k\times X}$

1.4 $K_{N^2\times X} \leftarrow (\mu\nu|P)$

1.5 $B_{N^2\times k} \leftarrow K_{N^2\times X}\Omega_{X\times k}$

Step 2. Form $\tilde{B}^{W}$ for the SVD, NAF, STO RI approximations

2.1 $B_{N^2\times k} = U_{N^2\times k/}S_{k/\times k/}L^{\dagger}_{k/\times k}$

2.2 (SVD) $\tilde{B}^{SVD}_{N^2\times\tilde{k}} = U_{N^2\times\tilde{k}}S_{\tilde{k}\times\tilde{k}}$

2.3 (NAF) $\tilde{B}^{NAF}_{N^2\times\tilde{k}} = B_{N^2\times k}L_{k\times\tilde{k}}$

2.4 (STO) Form $\Theta_{\tilde{k}\times k}$ and $\tilde{B}^{STO}_{N^2\times\tilde{k}} = B_{N^2\times k}\Theta^{\dagger}_{k\times\tilde{k}}$

2.5 $B_{VA\times k} \leftarrow C^{\dagger}_{V\times N}B_{N^2\times k}C_{N\times A}$

2.6 $B^{W}_{VA\times\tilde{k}} \leftarrow C^{\dagger}_{V\times N}B^{W}_{N^2\times\tilde{k}}C_{N\times A}$

Step 3. Evaluate MP2 correlation energy

For i,j in the active occupied MOs {

3.1 (RI) $Q^{ij}_{VV} = B_{iV\times\tilde{k}}B^{\dagger}_{\tilde{k}\times Vj}$

3.2 (SVD, NAF, STO) $Q^{ij}_{VV} = \tilde{B}^{W}_{iV\times\tilde{k}}\tilde{B}^{W,\dagger}_{\tilde{k}\times Vj}$

3.3 $A^{ij}_{ab} = Q^{ij}_{ab}/(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b)$

3.4 $E^{(2)} \leftarrow Q^{ij}_{VV}(2A^{ij}_{VV} - A^{ij,\dagger}_{VV})$}

2.4 Results and discussion

2.4.1 Correlation Energy.

The correlation energy predicted by the SVD RI-MP2 method is compared with the conventional RI-MP2 and/or full MP2 correlation energy of water clusters that contain 10-35 molecules. The cc-pVDZ[26] and cc-pVDZ-RI[27,28] basis sets, denoted cc-pVDZ//cc-pVDZ-RI, are used for the AO and the auxiliary bases, respectively. The water cluster geometries were randomly generated and were then optimized using the HF/3-21G* level of theory.[29] Since the RI-MP2/cc-pVDZ//cc-pVDZ-RI and the full MP2/cc-pVDZ results are almost identical (e.g., the average correlation energy difference is about 0.168 ± 0.085 kcal/mol), in the following, the SVD RI-MP2 correlation energy will only be discussed in terms of the RI-MP2 reference. The SVD RI-MP2 correlation energy error (ΔE_{RI}^{SVD}) is defined as the absolute difference between the correlation energy of water clusters calculated by the SVD RI-MP2 method $\left(E_{SVD}^{(2)}\right)$ and the correlation energy calculated using the standard RI-MP2/cc-pVDZ//cc-pVDZ-RI method $\left(E_{RI}^{(2)}\right)$.

$$\Delta E_{RI}^{SVD}\left(kcal/mol\right) = \left|E_{SVD}^{(2)} - E_{RI}^{(2)}\right| \tag{41}$$

As shown in FIGURE 2.2, the absolute error varies monotonically with respect to the SVD threshold θ_{SVD} or with the the compression factor $k/\tilde{k}$; i.e., the error increases when the SVD threshold or the compression factor increases. As the SVD threshold approaches zero or the compression factor approaches unity, the absolute error approaches zero as expected.

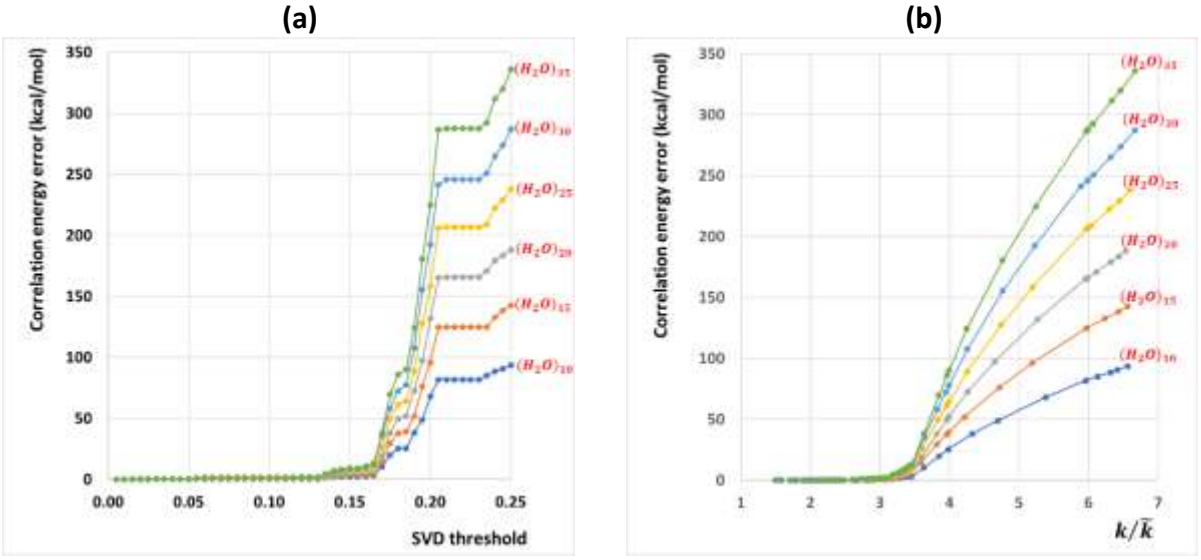


FIGURE 2.2 *The variation of the absolute error in the correlation energy for water clusters with respect to the SVD threshold (a) and the compression factor (b).*

The stepwise shape of FIGURE 2.2a vs. the smooth variation in FIGURE 2.2b occurs because the SVD threshold and the compression factor affect the SVD dimension slightly differently. On the one hand, any change in the compression factor $k/\tilde{k}$ modifies the dimensions of the SVD matrices (Eqs. (24)-(26)). On the other hand, it might require several incremental steps (e.g., of 0.005 in water cluster calculations) of the SVD threshold θ_{SVD} to hop from one singular value to another to actually truncate the set of SVD matrices.

It can also be seen from FIGURE 2.2a-b that the absolute error is size dependent; i.e., the error increases as the size of the water cluster increases. However, the change in the absolute error *per* correlated electron has the same trend and range for all water clusters as shown in FIGURE 2.3. The stochastic RI-MP2 correlation energy error per electron is also plotted in FIGURE 2.3 for comparison. For a low compression factor $k/\tilde{k}$ (e.g., below 4.0), the SVD RI-MP2 approach yields a lower error/electron than does the stochastic RI-MP2 approach, while the stochastic RI-

MP2 error/electron is *statistically* smaller for large compression factor (e.g., above 4.0). This is consistent with the observation by Takeshita *et* al.[21] that the stochastic RI-MP2 error does not change much as a function of the number of stochastic orbitals.

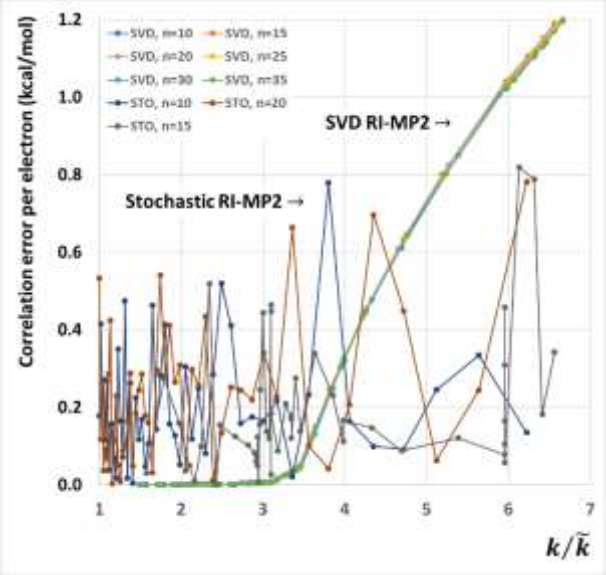


FIGURE 2.3 *The variation in the absolute error/electron in the correlation energy for water clusters calculated by the SVD and STO RI-MP2 methods. For clarity, for the STO RI-MP2 calculations, only three water clusters with the number of molecules n=10, 15, 20 were plotted.*

While the SVD RI-MP2 correlation energy error of water clusters increases and surpasses the error of the stochastic RI approximation as the compression factor increases (FIGURE 2.3), the SVD RI-MP2 error variation is monotonic and deterministic, in contrast to the stochastic approach. This is an indication that the SVD RI-MP2 method causes a *systematic* error, which could cancel out for the prediction of relative properties.

2.4.2 Potential energy surfaces.

An important relative property in chemistry is the potential energy surface (PES), which determines the equilibrium geometry and most dynamic properties of molecular systems. In this section, a part of the T-shape benzene dimer[30] PES obtained by varying the distance between the centers of two benzene monomers from 1.995–4.395 Å with an incremental step of 0.1 Å are numerically calculated by the SVD RI-MP2 method with low (e.g. 4.0) and high (e.g., 8.0) compression factors. The NAF and stochastic RI-MP2 methods are also used for comparison. The calculated results are validated against the conventional RI-MP2 PES. All calculations were done using the cc-pVDZ//cc-pVDZ-RI basis set.

To place the calculated results on the same scale, for each method, the energy of the T-shaped benzene dimer at the smallest distance of 1.995 Å is used as the reference to define the relative energy ΔE_W^f, in which W can be SVD, NAF or the STO RI-MP2 method and f is the compression factor of 4.0 or 8.0. For instance, the relative energy of the benzene dimer at distance d_i calculated by the SVD RI-MP2 with the compression factor 4.0 is defined as

$$\Delta E_{SVD}^{4.0}\Big|_{d=d_i} = E_{SVD}^{4.0}\Big|_{d=d_i} - E_{SVD}^{4.0}\Big|_{d=1.995\text{Å}} \tag{42}$$

The relative energy for the conventional RI-MP2 calculations ΔE_{RI} can also be defined in the same way. Plots of the SVD, NAF, STO RI-MP2 and the conventional RI-MP2 relative energies for the T-shaped benzene dimer distance are shown in FIGURE 2.4a-b. The conventional RI-MP2 PES is almost exactly reproduced by the SVD RI-MP2 method for both compression factors of 4.0 and 8.0, whereas the stochastic and the NAF RI-MP2 methods show very little explicit correlation with

the reference RI-MP2 PES. Although the NAF RI-MP2 PES does not fluctuate as vigorously as the STO RI-MP2 PES, its shape does reflect the non-deterministic nature as discussed earlier.

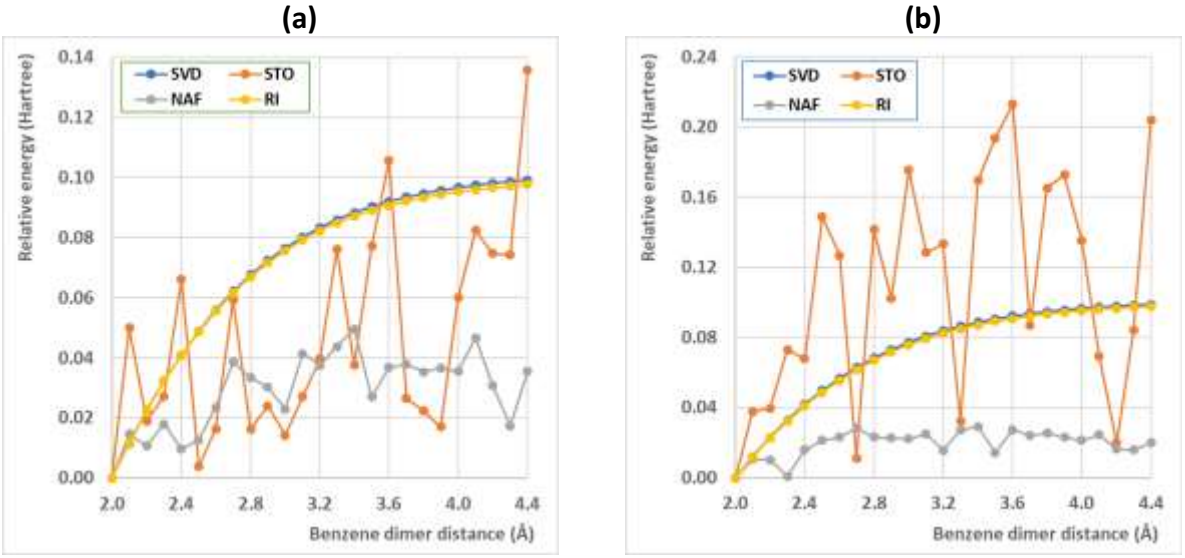


FIGURE 2.4 *Part of the potential energy surface of the T-shape benzene dimer calculated by the SVD RI-MP2, NAF RI-MP2, STO RI-MP2 and conventional RI-MP2 methods with compression factors of (a) 4.0 and (b) 8.0.*

The correlation between the approximate PESs and the conventional RI-MP2 PES can be assessed using the least squares regression method.[31] In the least squares regression method, the relation between two data sets is represented by a least squares regression equation (also called a least squares fitting equation) that is obtained by minimizing the squares of the offset of data points from the fitting equation. A popular relation between data sets is the linear regression. An index is also introduced to quantitatively assess the correlation between two data sets, called the (linear) cross-correlation coefficient.[32] The cross-correlation coefficient R^2 can vary from 0.0 for no correlation between data sets to 1.0 for perfectly correlated data sets. Perfectly correlated data sets means that if one data set is known, the other data set can be

calculated from the least squares regression equation.[31] The (linear) cross-correlation coefficient between the SVD, NAF and STO RI-MP2 PESs with the conventional RI-MP2 PES can be formulated as follows:

$$R^2 = \frac{\left[\sum_i \left(\Delta E_W^f\big|_{d=d_i} - \overline{\Delta E}_W^f\right)\left(\Delta E_{RI}^f\big|_{d=d_i} - \overline{\Delta E}_{RI}\right)\right]^2}{\sum_i \left(\Delta E_W^f\big|_{d=d_i} - \overline{\Delta E}_W^f\right)^2 \sum_i \left(\Delta E_{RI}^f\big|_{d=d_i} - \overline{\Delta E}_{RI}\right)^2} \tag{43}$$

In Eq. (43) $\overline{\Delta E}_W^f$ is the mean of the relative energy ΔE_W^f (c.f. Eq. (42)) calculated by the method *W* (SVD, NAF or STO RI-MP2); $\overline{\Delta E}_{RI}$ is the mean of the relative energy ΔE_{RI} calculated by the conventional RI-MP2 method.

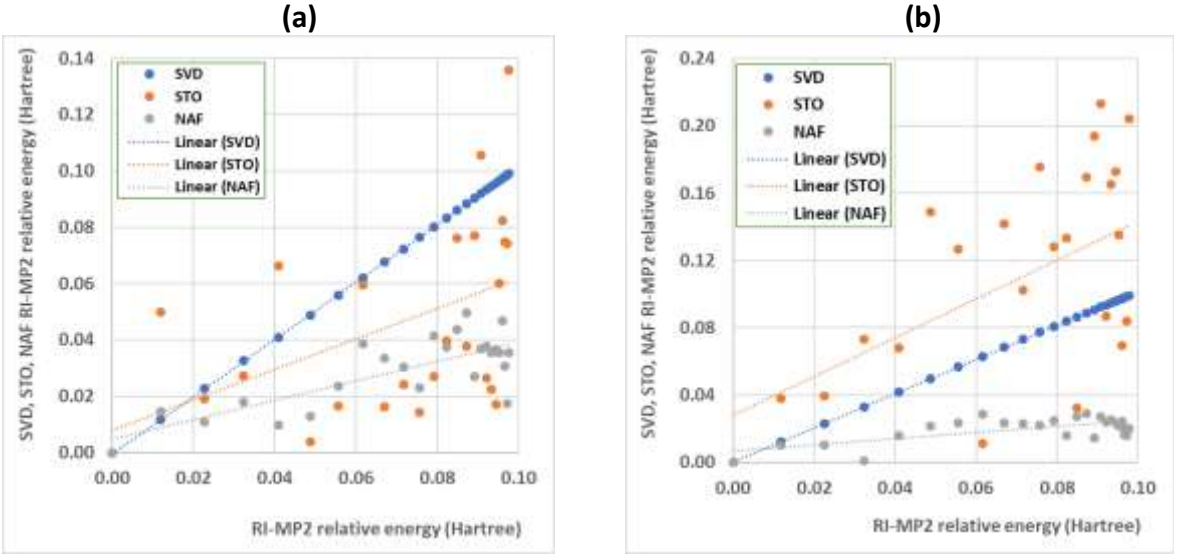


FIGURE 2.5 *Linear correlation between the SVD, NAF and stochastic RI-MP2 methods compared to the conventional RI-MP2 PES with compression factors of 4.0 (a) and 8.0 (b).*

To predict the relation between the SVD, NAF and STO RI-MP2 PESs with the reference RI-MP2 PES, the relative energy ΔE_W^f (cf. Eq. (42)) is plotted in FIGURE 2.5 against the conventional RI-MP2 relative energy ΔE_{RI}. It can be seen that the SVD RI-MP2 data points (blue dots) coincide

with the blue lines that represented the linear regression Eqs. (44) and (45). The corresponding cross-correlation coefficients R^2 are ~1.0, implying a strong correlation between the SVD RI-MP2 and conventional RI-MP2 methods:

$$\Delta E_{SVD}^{4.0} = 1.0181 \times \Delta E_{RI} - 0.0005; \qquad R^2 = 1.0000 \tag{44}$$

$$\Delta E_{SVD}^{8.0} = 1.0124 \times \Delta E_{RI} + 0.0003; \qquad R^2 = 0.9999 \tag{45}$$

On the other hand, the NAF (grey dots) and stochastic (orange dots) RI-MP2 data points are scattered throughout the variable space with no apparent explicit correlation. An effort to fit these data points to the linear regression equations (46)-(49), represented by the grey and orange dotted lines in FIGURE 2.5a-b, results in very small cross-correlation coefficients. For instance, the largest cross-correlation coefficient for the NAF RI-MP2 method with a compression factor of 4.0 is only ~0.6. The other R^2 values for NAF RI-MP2 and STO RI-MP2 are much smaller. The cross correlation coefficients also show that the NAF method performs better than the STO RI approximation.

$$\Delta E_{NAF}^{4.0} = 0.3445 \times \Delta E_{RI} + 0.0048; \qquad R^2 = 0.6069 \tag{46}$$

$$\Delta E_{NAF}^{8.0} = 0.1804 \times \Delta E_{RI} + 0.0068; \qquad R^2 = 0.457 \tag{47}$$

$$\Delta E_{STO}^{4.0} = 0.5423 \times \Delta E_{RI} + 0.0079; \qquad R^2 = 0.2152 \tag{48}$$

$$\Delta E_{STO}^{8.0} = 1.1509 \times \Delta E_{RI} + 0.0281; \qquad R^2 = 0.2744 \tag{49}$$

2.4.3 Binding energies.

Now, consider the binding energies of non-covalent complexes. A database of 22 non-covalent complexes called the S22 test set[24] has been established to validate electronic structure methods. In this section, the SVD, NAF and STO RI-MP2 methods are used to calculate the binding energies of the S22 non-covalent complexes with the compression factor $k/\tilde{k}$ varying from a moderate to large value (i.e., 5.0-20.0). The calculated results are validated against the conventional RI-MP2 and/or full MP2 methods. All calculations were done using the cc-pVTZ/cc-pVTZ-RI basis set.

Since the conventional RI-MP2 and full MP2 results are almost identical (e.g., the average binding energy difference is about 0.029 ± 0.091 kcal/mol), the conventional RI-MP2 results are used as the reference in the following discussion. The binding energy along with the average error relative to the RI-MP2 values ($\overline{\Delta E}^W$, see Eq. (50)), and the error radius (σ, see Eq. (51)) are shown in TABLE 1. The average error of the method *W* (SVD, NAF or STO RI-MP2) is defined as the average of the absolute difference between the binding energy calculated by the method *W* and that of the RI-MP2 reference.

$$\overline{\Delta E}^W = \frac{1}{22}\sum_{i=1}^{22}\left|E_i^W - E_i^{RI}\right| \tag{50}$$

Here, E_i^W and E_i^{RI} are the binding energy of the i^{th} complex calculated by the method *W* and by the reference RI-MP2 method, respectively. The error radius of the binding energy, which is also the standard deviation, can be formulated as

$$s^W = \sqrt{\frac{1}{22}\sum_{i=1}^{22}\left(\left|E_i^W - E_i^{RI}\right| - \overline{\Delta E}^W\right)^2} \quad (51)$$

TABLE 1 shows that both the average error and the error radius of the SVD RI-MP2 method are ~1.0 kcal/mol; those of the NAF and STO RI-MP2 methods are about 10-100 kcal/mol. Specifically, for the SVD RI-MP2 method, the average absolute error is below 1.0 kcal/mol for compression factors up to 15.0; when the compression factor goes up to 20.0, the average error remains below 1.5 kcal/mol. Both the NAF and STO RI-MP2 methods fail to estimate the binding energies of the S22 test set even qualitatively correctly.

While the binding energy is well reproduced by the SVD approximation for a wide range of compression factors, as demonstrated for water clusters, the absolute SVD energy error (e.g., of the bound state complex) compared to the regular RI reference can be large. For instance, the absolute SVD energy error is about 25.9-144.7 kcal/mol for compression factors varying from 1.0-20.0. The corresponding STO absolute energy error is about 209.9-1000.4 kcal/mol. Therefore, it is likely that the STO approximation introduces relatively poor binding energies compared with the SVD approximation as shown in TABLE 1. Interestingly, in some cases, the SVD absolute energy error can be much larger than that of the STO approximation, but the SVD binding energy error remains very small compared with STO approximation. For instance, for water dimer, the SVD and STO energy error [compared with the regular RI reference] are 12.3 and 1.6 kcal/mol, respectively; the SVD and STO *binding* energy errors [see row 15th in TABLE 1] are 0.1 and 37.9 kcal/mol, respectively. These results reflect the precision characteristics of the SVD approximation, which will be further discussed in the next section.

TABLE 2.1 Binding energies (kcal/mol) of the S22 test set calculated by the SVD RI-MP2, NAF RI-MP2, STO RI MP2 and RI-MP2 methods using the cc-pVTZ//cc-pVTZ-RI basis set.

Method	**SVD RI-MP2**				**NAF RI-MP2**				**Stochastic RI-MP2**				**RI-MP2**
$k/\tilde{k}$	**5.0**	**10.0**	**15.0**	**20.0**	**5.0**	**10.0**	**15.0**	**20.0**	**5.0**	**10.0**	**15.0**	**20.0**	
Uracil dimer h-bonded	-20.0	-20.0	-20.0	-25.7	-8.2	-19.3	-8.2	-8.2	252.3	264.5	989.4	-43.8	-20.1
Formic acid dimer	-18.9	-21.9	-17.7	-20.0	-25.6	-9.5	-25.6	-25.6	258.1	138.0	511.3	-3.0	-18.9
2-pyridoxine 2-aminopyridine	-17.7	-17.5	-18.1	-18.1	-20.3	-16.8	-20.3	-20.3	417.2	-67.8	-431.0	-57.5	-17.2
Adenine thymine Watson-Crick	-17.3	-17.2	-17.8	-18.5	-7.4	-5.6	-7.4	-7.4	20.7	29.1	735.1	818.8	-16.9
Formamide dimer	-15.7	-15.6	-16.7	-15.7	-0.2	-10.2	-0.2	-0.2	-386.3	25.3	153.1	14.5	-15.6
Adenine thymine complex stack	-15.7	-15.3	-16.3	-16.5	-4.7	-3.9	-4.7	-4.7	60.3	-123.8	-929.1	-113.7	-15.3
Uracil dimer stack	-10.6	-9.9	-10.5	-13.3	1.5	-0.5	1.5	1.5	270.2	100.9	-171.6	378.2	-10.2
Indole benzene complex stack	-7.9	-7.9	-9.2	-8.7	-2.7	3.1	-2.7	-2.7	-966.0	147.0	50.5	-53.1	-7.8
Phenol dimer	-8.3	-8.2	-8.6	-8.6	-59.1	-35.1	-59.1	-59.1	-291.3	696.9	446.0	35.3	-8.1
Indole benzene T-shape complex	-6.7	-6.8	-7.2	-7.0	-5.1	-0.6	-5.1	-5.1	-142.0	-1372.7	-526.6	358.6	-6.6
Pyrazine dimer	-6.3	-6.5	-6.4	-9.4	3.0	-1.5	3.0	3.0	-47.2	25.2	-896.3	-347.3	-6.2
Benzene HCN complex	-4.5	-5.4	-4.8	-4.8	-17.0	-16.7	-17.0	-17.0	-68.1	-42.6	51.6	323.7	-4.4
Water dimer	-5.2	-5.4	-8.7	-6.5	-7.8	-0.9	-7.8	-7.8	-43.2	-74.3	38.9	18.4	-5.3
Benzene dimer parallel displaced	-4.6	-4.8	-5.4	-5.5	-8.0	2.2	-8.0	-8.0	580.3	-41.1	-88.3	285.4	-4.6
Benzene dimer T-shaped	-3.5	-3.7	-3.8	-4.2	19.0	3.1	19.0	19.0	-30.0	681.7	-144.6	-31.7	-3.5
Benzene water complex	-3.7	-3.6	-3.7	-3.9	4.0	4.0	4.0	4.0	515.1	54.4	-8.5	-110.2	-3.7
Ammonia dimer	-3.4	-3.4	-4.6	-3.6	-6.0	5.3	-6.0	-6.0	-5.3	69.0	109.6	-100.9	-3.4
Benzene ammonia complex	-2.4	-2.5	-2.6	-2.9	10.1	2.2	10.1	10.1	382.9	117.4	-26.1	-19.3	-2.4
Benzene - Methane complex	-1.7	-1.8	-1.8	-1.9	-36.3	-19.8	-36.3	-36.3	67.0	-375.2	-313.8	-24.6	-1.7
Ethene ethyne complex	-1.4	-1.6	-1.4	-1.4	1.8	-3.3	1.8	1.8	-44.0	126.0	76.1	-9.2	-1.5
Ethene dimer	-1.4	-1.4	-1.1	-1.2	-18.9	-14.6	-18.9	-18.9	19.0	127.1	-32.9	-0.6	-1.4
Methane dimer	-0.5	-0.3	-0.3	-2.4	6.9	2.6	6.9	6.9	-161.8	34.9	20.4	15.1	-0.4
$\overline{\Delta E^W}$	**0.12**	**0.29**	**0.67**	**1.14**	**11.91**	**8.38**	**11.91**	**11.91**	**230.82**	**218.2**	**308.2**	**143.90**	
σ^W	**0.14**	**0.62**	**0.73**	**1.29**	**11.35**	**5.88**	**11.35**	**11.35**	**234.33**	**312.66**	**320.46**	**143.88**	

2.4.4 SVD accuracy and precision.

While the SVD RI approximation applied to the MP2 method yields increasingly large errors for absolute properties (e.g., the correlation energy) when increasing the compression factor, its relative properties (e.g., PES and binding energy) retain reasonable accuracy. The origin of this behavior is very likely a systematic error cancellation due to the precision characteristics of the approximation.[33] Using the S22 binding energy calculations, the accuracy and precision of the SVD, NAF and STO RI approximations can be directly examined using the Frobenius norm (*F*) and the relative Frobenius norm (ΔF) of the 4-2ERI matrix defined as follows:

$$\Delta F = \left| F_b^W - F_u^W \right| \tag{52}$$

For simplicity, the matrix (VA|VA) with V=virtual, A=active occupied MOs, which is the main ingredient for the MP2 correlation energy, is used. F_b^W is the Frobenius norm[20] of the difference of the matrix (VA|VA) evaluated by the approximation *W* (i.e., SVD, NAF and STO RI) from that of the regular RI approximation for an S22 complex in its bound state. Similarly, F_u^W is the Frobenius norm of the same S22 complex in its unbound state. The Frobenius norm (e.g., F_b^W) represents the absolute deviation [accuracy] of the approximation *W*, while the relative Frobenius ΔF describes the precision of the approximation.

FIGURE 2.6a shows that the SVD relative Frobenius norm for the H-bonded uracil dimer [the first complex in TABLE 1], for instance, remains small [e.g., 0-0.0028] when varying the compression factor from 1.0-20.0. This results in a relatively low SVD binding energy error as discussed in the previous section. Nevertheless, the small relative Frobenius norm alone does not guarantee the accuracy of the relative property [e.g., binding energy] as the error cancellation

becomes sensitive when the absolute error [Frobenius norm] gets large. For instance, the relative Frobenius norms for the H-bonded uracil dimer are the same for the compression factors of 5.0 and 20.0, about 0.0025. However, the binding energy error of the latter is larger (5.6 vs. 0.1 kcal/mol, see TABLE 1) due to the larger absolute error reflected through the Frobenius norm [FIGURE 2.6b].

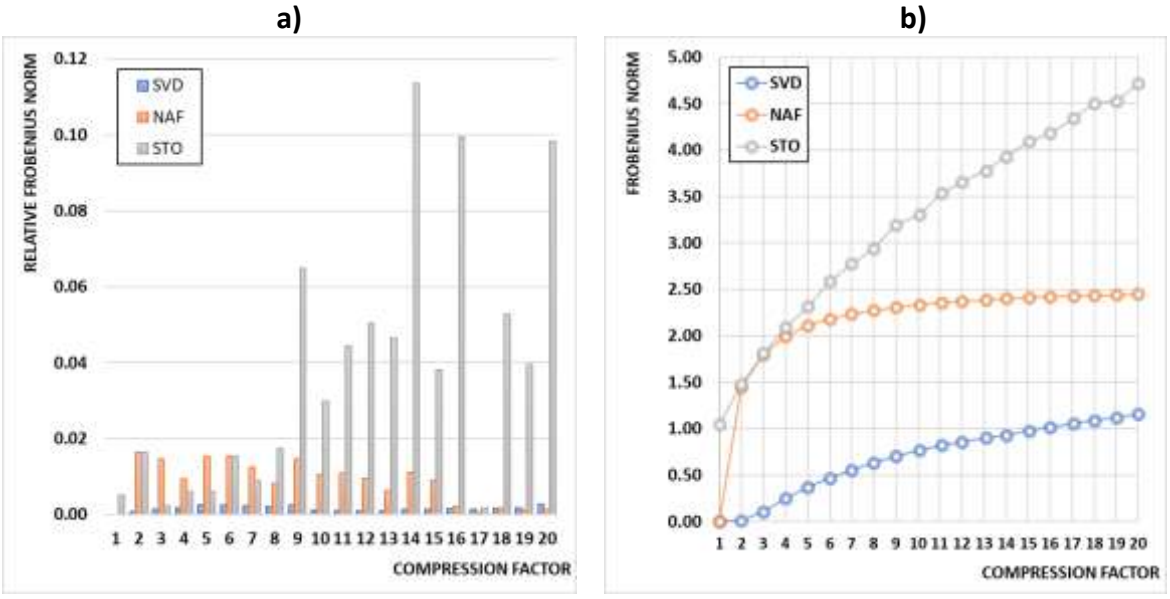


FIGURE 2.6 *Relative Frobenius norm (a) and Frobenius norm (b) of the (VA|VA) matrix from the H-bonded uracil dimer calculations using the SVD, NAF and STO approximation.*

2.4.5 Comment on computational efficiency

Since the SVD RI-MP2 code has not yet been fully optimized and parallelized, this section only briefly comments on the computational cost of the SVD RI-MP2 method relative to the standard RI-MP2 method[34] that has been implemented and optimized in GAMESS.[35] The RI-MP2 wall time for some S22 dimer calculations are recorded in TABLE 2. The calculations were done using a single core of a 96-cpu Intel(R) Xeon(R) Platinum 8160 CPU @2.10GHz chip. The Intel Fortran compiler and the MKL library were used for compiling and linking with the linear algebra library.

In TABLE 2, the SVD RI-MP2 wall time has been split into two parts. The first component is the time to execute the SVD denoted **DGESVD**. The other component includes time for the 2- and 3-2ERI evaluation and the AO to MO transformation, forming the 4-2ERI and MP2 amplitudes, and the MP2 correlation energy calculation. This part is denoted **CORR**.

The SVD RI-MP2 speedup Δt_{RI}^{SVD} (columns 4, 8 and 12 in TABLE 2) is defined as the ratio of the wall time of the standard RI-MP2 (t^{RI}) to the wall time of the SVD RI-MP2 calculation (t^{SVD}).

$$\Delta t_{RI}^{SVD} = \frac{t^{RI}}{t^{SVD}} \tag{53}$$

These times measure only the correlation energy (RI-MP2) part of the calculations.

The absolute binding energy errors $\Delta E_{RI}^{SVD} = \left| E_b^{SVD} - E_b^{RI} \right|$ (columns 5, 10 and 15), are also given in TABLE 2.

Compared with the conventional RI-MP2 method, the SVD RI-MP2 method introduces a **DGESVD** overhead, while the computational saving is gained in the **CORR** part. In total, the SVD RI-MP2 is about $1.6 - 2.7$ times faster than the conventional RI-MP2. The loss of binding energy accuracy is $\sim 0.1 - 1.6$ kcal/mol. For instance, for the adenine-thymine Watson-Crick complex, with compression factor $k/\tilde{k} = 5.0$, the speedup Δt_{RI}^{SVD} is 1.6 and the binding energy error is 0.4 kcal/mol. For the adenine-thymine π-stacked complex, with a large compression factor of 20.0, the speedup is 2.7 with a binding energy error below 1.0 kcal/mol.

Since the **DGESVD** is fixed for a given molecular system, more computational savings will be obtained when applying the SVD RI approximation to higher level *ab initio* methods (e.g., MP2 gradient, coupled cluster methods, larger basis sets). TABLE 2 also shows that future code

optimization can be focused on two parts: (1) optimization of the SVD routine and (2) optimization of the **CORR** part. In the **CORR** part, the compression of the fundamental RI matrix combined with a shared memory model, such as OpenMP,[36] can, for instance, enhance the on-node memory usage. Further code optimization will be addressed in a later work.

TABLE 2.2 Timing for the SVD RI-MP2 and standard RI-MP2 calculations for three S22 dimers. **DGESVD** is the SVD routine overhead, which is 0.0 (s) for standard RI-MP2 calculations. **CORR** includes the times for the 2- and 3-2ERI integral evaluations and the AO to MO 4-label transformation, forming 4-2ERI, MP2 amplitude and correlation energy evaluation. Δt_{RI}^{SVD} is the ratio of the SVD RI-MP2 wall time relative to the conventional RI-MP2. ΔE_{RI}^{SVD} is the absolute difference in the SVD RI-MP2 binding energy relative to the conventional RI-MP2 method.

	Adenine thymine Watson-Crick complex				**Adenine thymine complex stack**				**Indole benzene complex stack**			
$\boldsymbol{k/\tilde{k}}$	**DGESVD**	**CORR**	Δt_{RI}^{SVD}	ΔE_{RI}^{SVD}	**DGESVD**	**CORR**	Δt_{RI}^{SVD}	ΔE_{RI}^{SVD}	**DGESVD**	**CORR**	Δt_{RI}^{SVD}	ΔE_{RI}^{SVD}
5.0	14.3	19.4	1.6	0.4	7.8	7.7	1.9	0.1	14.1	19.7	1.6	0.4
10.0	14.4	10.9	2.1	0.3	7.9	4.2	2.5	0.1	14.3	11.0	2.2	0.0
15.0	14.6	8.4	2.3	0.9	7.8	3.5	2.7	1.4	14.2	8.1	2.5	1.0
20.0	15.0	8.1	2.3	1.6	7.8	3.4	2.7	0.9	14.9	8.2	2.4	1.2
RI-MP2	**0.0**	**52.6**	**1.0**	**0.0**	**0.0**	**30.1**	**1.0**	**0.0**	**0.0**	**55.2**	**1.0**	**0.0**

2.5. Concluding remarks

A compression scheme for the matrix of 4-2ERIs has been formulated by applying the SVD on top of the RI approximation. The accuracy (and computational cost) of the RI approximation can be controlled by varying the SVD singular value threshold or the compression factor. The application of the SVD RI approximation to the closed shell MP2 method shows that while errors in absolute energies (e.g., correlation energies of water clusters) can vary from small to large and are size dependent, the relative properties (e.g., the potential energy surface of the T-shape benzene dimer, the binding energy of the S22 non-covalent complexes) are accurately

reproduced. The accuracy of the SVD RI-MP2 method is much better than that of either the NAF approach or the STO approach. As the SVD RI approximation maintains the symmetric form of the RI approximation, the approximation is, in principle, readily applicable to DFT and/or other *ab initio* methods (e.g., couple cluster) that are compatible with the RI approximation.

Acknowledgements. This work was supported by a grant from the Department of Energy Exascale Computing Project (ECP), administered by the Ames Laboratory. The authors gratefully acknowledge many stimulating and insightful discussions with Mr. Viet Tran and Dr. Luke Roskop.

References

(1) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(2) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of Approximate Integrals in Ab Initio Theory. An Application in MP2 Energy Calculations. *Chem. Phys. Lett.* **1993**, *208*, 359–363.

(3) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral Approximations for LCAO-SCF Calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(4) Schmitz, G.; Madsen, N. K.; Christiansen, O. Atomic-Batched Tensor Decomposed Two-Electron Repulsion Integrals. *J. Chem. Phys.* **2017**, *146*, 134112.

(5) Hohenstein, E. G.; Parrish, R. M.; Sherrill, C. D.; Martínez, T. J. Communication: Tensor Hypercontraction. III. Least-Squares Tensor Hypercontraction for the Determination of Correlated Wavefunctions. *J. Chem. Phys.* **2012**, *137*, 221101.

(6) Parrish, R. M.; Hohenstein, E. G.; Martínez, T. J.; Sherrill, C. D. Tensor Hypercontraction. II. Least-Squares Renormalization. *J. Chem. Phys.* **2012**, *137*, 224106.

(7) Hohenstein, E. G.; Parrish, R. M.; Martínez, T. J. Tensor Hypercontraction Density Fitting. I. Quartic Scaling Second- and Third-Order Møller-Plesset Perturbation Theory. *J. Chem. Phys.* **2012**, *137*, 44103.

(8) Hummel, F.; Tsatsoulis, T.; Grüneis, A. Low Rank Factorization of the Coulomb Integrals for Periodic Coupled Cluster Theory. *J. Chem. Phys.* **2017**, *146*, 124105.

(9) Aquilante, F.; Delcey, M. G.; Pedersen, T. B.; Fdez. Galván, I.; Lindh, R. Inner Projection

Techniques for the Low-Cost Handling of Two-Electron Integrals in Quantum Chemistry. *Mol. Phys.* **2017**, *115*, 2052–2064.

(10) Røeggen, I.; Wisløff-Nilssen, E. On the Beebe-Linderberg Two-Electron Integral Approximation. *Chem. Phys. Lett.* **1986**, *132*, 154–160.

(11) Røeggen, I.; Johansen, T. Cholesky Decomposition of the Two-Electron Integral Matrix in Electronic Structure Calculations. *J. Chem. Phys.* **2008**, *128*, 194107.

(12) Koch, H.; Sánchez de Merás, A.; Pedersen, T. B. Reduced Scaling in Electronic Structure Calculations Using Cholesky Decompositions. *J. Chem. Phys.* **2003**, *118*, 9481–9484.

(13) Löwdin, P. -O. Some Properties of Inner Projections. *Int. J. Quantum Chem.* **1970**, *5*, 231–237.

(14) Beebe, N. H. F.; Linderberg, J. Simplifications in the Generation and Transformation of Two-electron Integrals in Molecular Calculations. *Int. J. Quantum Chem.* **1977**.

(15) Aquilante, F.; Lindh, R.; Pedersen, T. B. Analytic Derivatives for the Cholesky Representation of the Two-Electron Integrals. *J. Chem. Phys.* **2008**, *129*, 034106.

(16) Pedersen, T. B.; Aquilante, F.; Lindh, R. Density Fitting with Auxiliary Basis Sets from Cholesky Decompositions. *Theor. Chem. Acc.* **2009**, *124*, 1–10.

(17) Hohenstein, E. G.; Sherrill, C. D. Density Fitting and Cholesky Decomposition Approximations in Symmetry-Adapted Perturbation Theory: Implementation and Application to Probe the Nature of π-π Interactions in Linear Acenes. *J. Chem. Phys.* **2010**, *132*, 184111.

(18) Peng, B.; Kowalski, K. Highly Efficient and Scalable Compound Decomposition of Two-Electron Integral Tensor and Its Application in Coupled Cluster Calculations. *J. Chem. Theory Comput.* **2017**, *13*, 4179–4192.

(19) Peng, B.; Kowalski, K. Low-Rank Factorization of Electron Integral Tensors and Its Application in Electronic Structure Theory. *Chem. Phys. Lett.* **2017**, *672*, 47–53.

(20) Weisstein, E. W. Frobenius Norm. http://mathworld.wolfram.com/FrobeniusNorm.html.

(21) Takeshita, T. Y.; de Jong, W. A.; Neuhauser, D.; Baer, R.; Rabani, E. Stochastic Formulation of the Resolution of Identity: Application to Second Order Møller–Plesset Perturbation Theory. *J. Chem. Theory Comput.* **2017**, *13*, 4605–4610.

(22) Kállay, M. A Systematic Way for the Cost Reduction of Density Fitting Methods. *J. Chem. Phys.* **2014**, *141*, 244113.

(23) Weisstein, E. W. Gram-Schmidt Orthonormalization.

(24) Rezac, J.; Jurecka, P.; Riley, K.; Černý, J.; Valdés, H.; Pluháčková, K.; Berka, K.; Řezáč, T.; Pitoňák, M.; Vondrasek, J.; et al. Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (Www.Begdb.Com): A Users Manual and Examples. *Collect. Czechoslov. Chem. Commun.* **2008**, *73*, 1261–1270.

(25) The General Atomic and Molecular Electronic Structure System (GAMESS 2018-R3). https://www.msg.chem.iastate.edu/gamess/.

(26) Dunning, 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.

(27) Hättig, C. Optimization of Auxiliary Basis Sets for RI-MP2 and RI-CC2 Calculations: Core–Valence and Quintuple-ζ Basis Sets for H to Ar and QZVPP Basis Sets for Li to Kr. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.

(28) Weigend, F.; Köhn, A.; Hättig, C. Efficient Use of the Correlation Consistent Basis Sets in Resolution of the Identity MP2 Calculations. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

(29) Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. Self-Consistent Molecular-Orbital Methods. 22. Small Split-Valence Basis Sets for Second-Row Elements. *J. Am. Chem. Soc.* **1982**, *104*, 2797–2803.

(30) DiStasio, R. A.; von Helden, G.; Steele, R. P.; Head-Gordon, M. On the T-Shaped Structures of the Benzene Dimer. *Chem. Phys. Lett.* **2007**, *437*, 277–283.

(31) Weisstein, E. W. Least Squares Fitting.

(32) Weisstein, E. W. Correlation Coefficient.

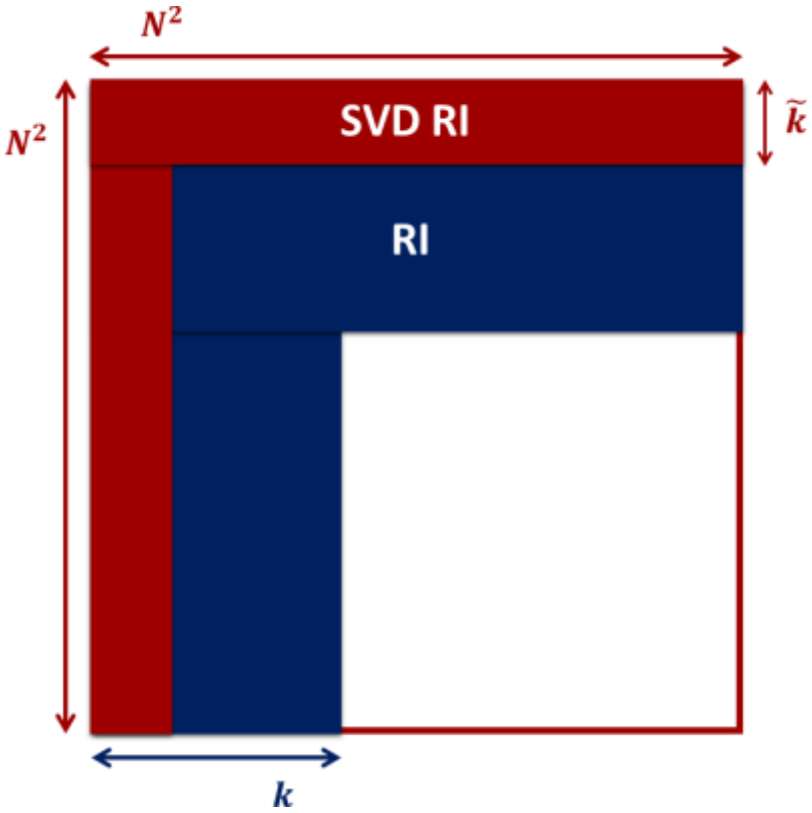
(33) Rebolini, E.; Izsák, R.; Reine, S. S.; Helgaker, T.; Pedersen, T. B. Comparison of Three Efficient Approximate Exact-Exchange Algorithms: The Chain-of-Spheres Algorithm, Pair-Atomic Resolution-of-the-Identity Method, and Auxiliary Density Matrix Method. *J. Chem. Theory Comput.* **2016**, *12*, 3514–3522.

(34) Katouda, M.; Nagase, S. Efficient Parallel Algorithm of Second-order Møller–Plesset Perturbation Theory with Resolution-of-identity Approximation (RI-MP2). *Int. J. Quantum Chem.* **2009**, *109*, 2121–2130.

(35) Gordon, M. S.; Schmidt, M. W. Chapter 41 - Advances in Electronic Structure Theory: GAMESS a Decade Later.; Frenking, G., Kim, K. S., Scuseria, G. E. B. T.-T. and A. of C. C., Eds.; Elsevier: Amsterdam, 2005; pp 1167–1189.

(36) OpenMP. https://www.openmp.org/.

Graphical abstract



CHAPTER 3. A HYBRID DISTRIBUTED/SHARED MEMORY MODEL FOR THE RI-MP2 METHOD IN THE FRAGMENT MOLECULAR ORBITAL FRAMEWORK

A paper submitted for publication at a later date

Buu Q. Pham and Mark S. Gordon

Abstract

The general distributed data interface (GDDI) that was developed for the fragment molecular orbital (FMO) method is combined with the shared memory OpenMP parallel middleware to support a threading multi-level parallelism. First, GDDI partitions [logical] compute nodes into groups, which are statically or dynamically assigned to different fragments. A small number of processes are created on each compute node. Each process subsequently spawns multiple threads for the actual computation. The performance of the hybrid GDDI/OpenMP approach relative to the pure GDDI model was examined in terms of the FMO/RI-MP2 method; i.e., the second-order Moller-Plesset (MP2) correlation energy was evaluated using the resolution-of-the-identity (RI) and the FMO approximations. The GDDI and OpenMP workload balances are handled by an arithmetic progression and a loop fusion, respectively. Other OpenMP properties, such as *threadprivate* or shared memory are combined with the low memory demand of the RI two-electron integrals to enhance the performance. Benchmark calculations demonstrate that because the hybrid parallel model can make use of multiprocessor resources more efficiently than the regular distributed memory-based GDDI model, calculations for small to large water clusters containing 139-2165 molecules exhibit speedups of a factor of 10x.

3.1 Introduction

The fragment molecular orbital (FMO) method[1–5] has been established as a computationally efficient approach to treat macromolecular systems with fully *ab initio* levels of accuracy. As the dispersion interaction usually plays an important role in many systems (e.g., water clusters, mesoporous silica nanoparticles, proteins), the dynamic correlation effect is usually introduced to the FMO framework using the second-order Moller-Plesset perturbation (MP2) method. Both the FMO/MP2 energy[6–9] and partially[10] and fully[11] analytic energy gradients are available. The FMO/MP2 method has also been combined with the-resolution-of-the-identity (RI) approximation[12,13] to reduce the cost of the integral transformation and memory storage. A version of the FMO/RI-MP2 energy[14] implemented in the GAMESS suite of programs[15,16] and gradient[17] in PAICS[18] were reported to be much faster than the FMO/MP2 method with a small loss of correlation energy and gradient accuracy.

As the FMO approach divides a molecule into fragments that can be treated essentially independently, it facilitates parallel implementation and reduces the (minimum) memory demand to that of the largest [monomer ($FMO^{(1)}$), dimer ($FMO^{(2)}$) or trimer ($FMO^{(3)}$)] fragment. In GAMESS, the MPI-based general distributed data interface (GDDI)[19] was designed to speed up FMO calculations. GDDI can divide *N* compute processes into *n* groups, such that each group contains *nprocs* processes [*nprocs* can vary among the groups] that can independently process fragments [chunks of work]. The fragment calculations can be statically or dynamically distributed among these groups of processes.

While the GDDI FMO implementation introduces linear scaling, its performance can be limited by the communication overhead of the distributed memory model. The recently popular

hybrid parallel model that combines the distributed model and the shared memory model (e.g., supported by an OpenMP API) can potentially speed up the calculations and reduce the memory footprint. In the GDDI/OpenMP model, a fragmentation calculation can be kicked off by a small number of processes (or ranks), split into a desired number of groups. Each rank then spawns a team of threads that perform the actual computation. The rank communication occurs through the expensive explicit message exchange [send/receive protocols], while the threads in a team can efficiently share data through the shared memory pool. Therefore, in most cases, to reduce the rank communication overhead, one rank is placed on a cpu socket, a numa (non-uniform memory access) node or the whole compute node depending on the computer architecture and code design. This rank then spawns a team of threads, which is usually equal to the number of physical or logical cores of the socket, numa node or compute node, respectively. Such a hybrid parallel model thereby introduces a benefit on clusters of large multicore compute nodes [e.g., Haswell, KNL], which is becoming the norm in most high-performance computing systems.

In this paper, the FMO/RI-MP2 method is implemented in terms of the hybrid distributed/shared memory parallel model [GDDI/OpenMP]. Its performance is examined relative to the pure GDDI FMO/RI-MP2 implementation.[14]

3.2 FMO/RI-MP2 energy

The FMO[1–5] and the FMO/RI-MP2 methods[14,17] have been well documented in the literature. The main idea of the FMO approximation is that while electrostatic interactions are delocalized, the exchange interaction and the induction are frequently localized. Therefore, a molecular system can (often based on chemical “intuition”) be partitioned into fragments so that *ab initio* calculations can be carried out for each fragment in the electrostatic potential of all other

fragments. The calculation is iterated until the charge distribution is self-consistent, called the self-consistent charge (SCC) procedure, to obtain the $FMO^{(1)}$ monomer energy. Two- (and possibly three-) body corrections are then performed to account for explicit fragment-fragment interactions, thereby introducing the $FMO^{(2)}$ and $FMO^{(3)}$ energies, respectively. For instance, the $FMO^{(1)}$ and $FMO^{(2)}$ energies are formulated as follows.

$$E = \underbrace{\underbrace{\sum_{I} E_I}_{FMO^{(1)}} + \sum_{I>J} \left(E_{IJ} - E_I - E_J \right)}_{FMO^{(2)}} \tag{1}$$

The fragment energy E_X can include the SCF (E_X^{SCF}) and the correlation energy E_X^{cor}. For instance, the RHF MP2 correlation energy $E_X^{(2)}$ is given by

$$E_X^{(2)} = \sum_{ij \in X}^{noccX} \sum_{ab \in X}^{virtX} A_{ab}^{ij,X} \left(2Q_{ab}^{ij,X} - Q_{ba}^{ij,X,\dagger} \right) \tag{2}$$

The indices *i,j* and *a,b* represent active occupied (*noccX*) and virtual (*virtX*) molecular orbitals (MOs), respectively. The dagger (†) indicates the matrix transpose operation (e.g., $Q_{ab}^{ij,X,\dagger} = Q_{ba}^{ij,X}$). $Q_{ab}^{ij,X}$ is the matrix of 4-index 2-electron repulsion integrals (4-2ERIs) transformed into the MO basis

$$Q_{ab}^{ij,X} = \sum_{\mu}^{aoX} C_{\mu i}^{X} \sum_{\nu}^{aoX} C_{\nu a}^{X} \sum_{\lambda}^{aoX} C_{\lambda j}^{X} \sum_{\sigma}^{aoX} (\mu\nu \mid \lambda\sigma) C_{\sigma b}^{X} \tag{3}$$

$$(\mu\nu \mid \lambda\sigma) = \iint dr_1 dr_2 \phi_\mu^*(r_1) \phi_\nu(r_1) r_{12}^{-1} \phi_\lambda^*(r_2) \phi_\sigma(r_2) \tag{4}$$

$C_{\mu p}^{X}$ is the p^{th} MO linear expansion coefficient in the AO basis $\{\phi_\mu\}$ of the fragment X. The matrix element $A_{ab}^{ij,X}$ is the ratio of $Q_{ab}^{ij,X}$ to the pairwise MO energy (ε_p^X) difference $D_{ab}^{ij,X}$:

$$A_{ab}^{ij,X} = Q_{ab}^{ij,X} / D_{ab}^{ij,X} \tag{5}$$

$$D_{ab}^{ij,X} = \varepsilon_i^X + \varepsilon_j^X - \varepsilon_a^X - \varepsilon_b^X \tag{6}$$

The MO transformation (Eq. (3)) and the correlation energy accumulation (Eq. (2)) are expensive steps. The MO transformation cost can be significantly improved by using the RI approximation,[12,20,21] in which a 4-2ERI is approximated by the product of 3- and 2-2ERIs.

$$(\mu\nu|\lambda\sigma) \simeq \sum_{PQ\in X}^{auxX} (\mu\nu|P) V_{PQ}^{X,-1} (Q|\lambda\sigma) \tag{7}$$

$$(mn|P) = \iint dr_1\, dr_2 f_m^*(r_1) f_n(r_1) r_{12}^{-1} a_P(r_2) \tag{8}$$

$$V_{PQ}^{X} = \iint dr_1 dr_2 \alpha_P(r_1) r_{12}^{-1} \alpha_Q(r_2) \tag{9}$$

In Eqs. (8) and (9), $\{\alpha_P\}$ or the index $\{P\}$ stands for the auxiliary basis functions of the fragment X (*auxX*). In GAMESS, the matrix V^X of 2-2ERIs is inverted and decomposed into the matrix Ω^X (Eq. (10)). The 3-2ERIs (Eq. (8)) are transformed into the MO basis (Eq. (11)), and then combined with Ω^X to form the fundamental RI matrix B^X (Eq. (12)), which can be stored in distributed arrays for further calculations; e.g., forming 4-2ERIs in the MO basis (Eq. (13)) .

$$V_{PQ}^{X,-1} = \sum_{R\in X}^{auxX} W_{PR}^{X} W_{RQ}^{X,\dagger} \tag{10}$$

$$\tilde{B}_{pq,P}^{X} = \sum_{\mu\in X}^{auxX} C_{\mu p}^{X,\dagger} \sum_{\nu\in X}^{auxX} (\mu\nu|P) C_{\nu q}^{X} \tag{11}$$

$$B^{X}_{pq,P} = \sum_{R\in X}^{auxX} \tilde{B}^{X}_{pq,R}\Omega^{X}_{RP} \tag{12}$$

$$\left(pq\,|\,rs\right) \simeq \sum_{R\in X}^{auxX} B^{X}_{pq,P}B^{X,\dagger}_{P,rs} \tag{13}$$

As the cost of forming the matrix B^X (Eq. (12)) in the parallel implementation is trivial compared with the correlation energy accumulation (Eq. (2) and (13)), the latter will be used to illustrate the GDDI/OpenMP FMO/RI-MP2 energy implementation. A serial *pseudocode* for this part of the algorithm is presented in SCHEME 3.1. As the rest of the paper mainly focuses on fragment parallelization, the following discussion addresses a particular monomer or dimer X [see Eq. (1)]. For simplicity, the superscript X will be dropped for the remainder of this work.

In SCHEME 3.1, first, the matrix B is formed and stored in the distributed memory. For each pair of active occupied MOs (i, j) [lines 4, 6], a portion of the matrix B is fetched to the process memory as the input for the for the $func\ corr$ [lines 12-15] to calculated the correlation energy contribution.

SCHEME 3.1. Serial RI-MP2 correlation energy implementation

```
1  do i = 1, nact
2        get B^i_DV
3        do j = 1, nact
4              get B^j_DV
5              E^(2) ← corr(i, j, B^i_DV, B^j_DV)
6        enddo j
7  enddo i
8  func corr(i, j, B^i_DV, B^j_DV)
9        Q^ij_VV = B^{i,†}_VD × B^j_DV
10       A^ij_VV = Q^ij_VV / D^ij_VV
11       Return A^ij_VV (Q^ij_VV + Q^ij_VV − Q^{ij,†}_VV)
```

3.3 GDDI/OpenMP FMO/RI-MP2 energy Implementation

3.3.1 GDDI/OpenMP model.

A normal GAMESS job using *N* nodes can kick off $N \times RPN$ processes, where RPN is the number of processes (ranks) *per* node. GDDI[19] can divide processes into groups of processes. These groups can be statically or dynamically assigned to fragment calculations (FIGURE 3.1a). As for DDI,[22] each GDDI process is accompanied by a dataserver to assist distributed arrays. The communication overhead and the memory footprint of the compute and dataserver processes, however, do not always allow GDDI to take full advantage of all processors in large multiproccessor compute nodes efficiently.

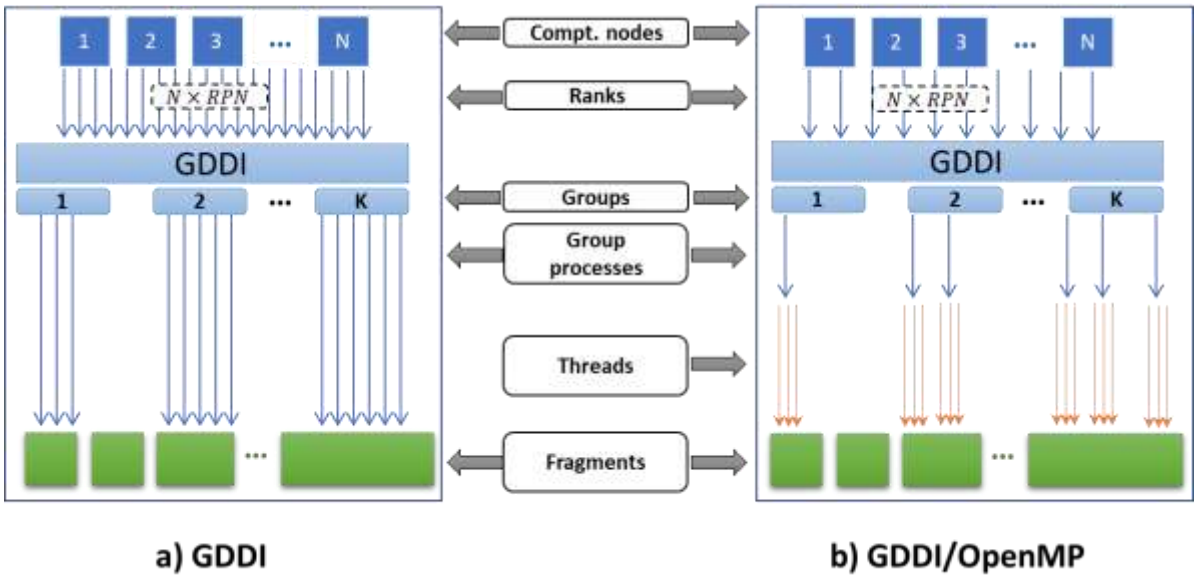


FIGURE 3.1 a) GDDI and b) hybrid GDDI/OpenMP models for the FMO method. *N* compute nodes generate $N \times RPN$ (ranks per node) processes, split into K groups and statically or dynamically assigned to a fragment computation. Fragment numbers depend on input parameters [e.g., point charge approximation] and fragment types [e.g., monomer, dimer or trimer]. The RPN for GDDI/OpenMP is made small, usually one rank per socket, numa node, or the whole compute node.

In the hybrid GDDI/OpenMP model, the RPN is made small (FIGURE 3.1b). Each rank then spawns a team of TPR (threads per rank) threads for actual computations. A small RPN introduces a small amount of replicated data, so large on-node shared data structures can be generated that subsequently reduce the data exchange. Therefore, the lightweight threads are likely to make use of the node resources (e.g., memory, physical cores) more efficiently, which will be demonstrated in the next sections.

3.3.2 GDDI load balance.

Each GDDI group is assigned to calculate a fragment correlation energy $E^{(2)}$, obtained by correlating $nocc \times nocc$ pairs of active occupied MOs (SCHEME 3.1) using virtual MOs. By using the index symmetry, the number of unique active occupied pairs can be reduced to $nocc \times (nocc + 1)/2$. This work can be distributed among $nprocs$ processes by assigning each process a chunk of the outermost loop $(Lstart, Lend)$ as shown in SCHEME 3.2 [line 2].

SCHEME 3.2 GDDI work distribution

```
1  //chunk of work for each rank
2  do i = Lstart, Lend
3        getB_DV^i
4        do j = 1, i
5              getB_DV^j
6              // index symmetry factor
7              f = 2.0
8              if(i = j) then f = 1.0
9              E^(2) ← f × corr(i, j, B_DV^i, B_VD^j)
10       enddo j
11 enddo i
```

The number of (i, j) pairs in each chunk of work is $\frac{(Lend-Lstart)(Lend-Lstart+1)}{2} + Lstart(Lend - Lstart)$ [lines 2, 4]. The ideal load balance can be achieved when the number of active occupied MOs is close to $nocc(nocc + 1)/(2 \times nprocs)$. This leads to a quadratic

recurrence relation (Eq. (14)) that can be used to retrieve all $(Lstart, Lend)$ starting from $Lstart = 1$ for process 0.

$$Lend^2 + Lend(1 + 2Lstart) + \left[\frac{nocc(nocc+1)}{nprocs} + Lstart - 3Lstart^2 \right] = 0 \qquad (14)$$

3.3.3 Shared memory data structure.

The RI 4-2ERIs are calculated by loading and combining the corresponding portions of the distributed matrix *B* (Eq. (13)) for each pair of active occupied MOs (i, j). The simplest approach is to get[22] B_{DV}^{i} and B_{DV}^{j} when needed as shown in SCHEMEs 3.1 and 3.2. Fetching distributed arrays, however, implies high communication overhead. The cost increases when one is repeatedly getting many small pieces of data, instead of a smaller number of large chunks of data. Due to the low memory footprint, the hybrid parallel model enables a large chunk of matrix *B* to be copied to the shared memory. This is further reinforced by the fact that, in the FMO framework, the calculations are on small fragments relative to the full system of interest. SCHEME 3.3 presents a simple algorithm to load as much of the matrix *B* as possible into the process memory [line 14]. When memory gets short, the replicated array is reduced, e.g., by 10% every repeating step.

SCHEME 3.3 Maximize on-node memory usage

```
1  //available memory per process
2  Pmem = totalmem/nprocs
3  nocc' = nocc
4  //simple loop to determine nocc'
5  ierr = 1
6  do while(ierr = 1)
7       Lmem = V × D × nocc'
8       if (lmem > pmem)then
9            //reduce memory by 10% if it does not fit
10           nocc' = 0.9 × nocc'
11           ierr = 1
12      else
13           //portion of B loaded into process memory
14           getB_DV^(1:nocc')
15           ierr = 0
16      endif
17 enddo
18 //calculate MP2 corr energy
19 do i = Lstart, Lend
20      //from B_DV^(1:nocc') or DDI mem
21      getB_DV^i
22      do j = 1, nocc
23           getB_DV^j
24           f = 2.0
25           if(i = j) then f = 1.0
26           E^(2) ← f × corr(i, j, B_DV^i, B_DV^j)
27      enddo j
28 enddo i
```

3.3.4 OpenMP load balance.

For clarity, SCHEME 3.3 is simplified and used as the starting point to illustrate the OpenMP[23] threading implementation. The chunk of GDDI work $(Lstart, Lend)$ can dynamically [e.g., using $schedule$ (DYNAMIC)][24] be assigned to threads by inserting OpenMP directives [SCHEME 3.4]. Each thread works on a set of (i, j) pairs and updates the correlation energy to the shared correlation energy variable $E^{(2)}$. To avoid race conditions, the shared variable is protected in a $critical$ region [lines 10-12]; i.e., only one thread at a time can update $E^{(2)}$, which is likely to spoil the performance.

SCHEME 3.4 Threading implementation

```
1  //OpenMP parallel region
2   !$OMP PARALLEL shared(E^(2))
3   //parallelize the loop with threads
4   !$omp do schedule(DYNAMIC)
5  do i = Lstart, Lend
6        do j = 1, nocc
7              f = 2.0
8              if(i = j) then f = 1.0
9   //using critical region to avoid race condition
10  !$omp critical
11             E^(2) ← f × corr(i, j, B_DV^i, B_DV^j)
12  !$omp end critical
13       enddo j
14  enddo i
15  !$OMP END PARALLEL
```

The race conditions and the *critical* region can be avoided by introducing a temporary *threadprivate* variable $\tilde{E}^{(2)}$ used to accumulate the correlation energy in each thread computation. When the loop finishes, $\tilde{E}^{(2)}$ is reduced to the shared variable $E^{(2)}$ using the efficient *atomic* write [lines 13-14] as shown in SCHEME 3.5.

SCHEME 3.5 Using threadprivate property

```
1  //define a threadprivate variable
2   !$omp threadprivate(Ẽ^(2))
3   !$OMP PARALLEL shared(E^(2))
4   !$omp do schedule(DYNAMIC)
5  do i = Lstart, Lend
6        do j = 1, nocc
7              f = 2.0
8              if(i = j) then f = 1.0
9              Ẽ^(2) ← f × corr(i, j, B_DV^i, B_DV^j)
10       enddo j
11 enddo i
12 //using atomic write
13  !$omp atomic
14       E^(2) ← Ẽ^(2)
15  !$OMP END PARALLEL
```

To remove the branching instruction and enhance the workload balance, the main loop in SCHEME 3.5 can be split into *two* loops [SCHEME 3.6]. The overhead due to splitting the loops is reduced by putting both loops in one OpenMP $PARALLEL$ region and using the $nowait$ clause [Lines 5, 12 in SCHEME 3.6].

SCHEME 3.6 Removing branch instruction

```
1  //define a threadprivate variable
2  !$omp threadprivate(Ẽ(2))
3  !$OMP PARALLEL shared(E(2))
4  //using nowait clause.
5  !$omp do nowait schedule(DYNAMIC)
6     do i = Lstart, Lend
7           do j = 1, i − 1
8                 Ẽ(2) ← 2.0 × corr(i, j, B_DV^i, B_DV^j)
9           enddo j
10    enddo i
11 //using nowait clause.
12 !$omp do nowait schedule(DYNAMIC)
13       do i = Lstart, Lend
14             Ẽ(2) ← 1.0 × corr(i, j, B_DV^i, B_DV^j)
15       enddo i
16 !$omp atomic
17       E(2) ← Ẽ(2)
18 !$OMP END PARALLEL
```

SCHEME 3.6 cannot completely remove the severe load *imbalance* of the triangular [trepazoidal] loop [lines 6, 7]. While the OpenMP API introduced the $collapse$ clause to fuse nested loops, it is only available for rectangular structures. For non-rectangular loops, a set of formulas has been suggested to fuse loops.[25] Another simple solution is to use a small buffer array tri of size $nocc \times (nocc - 1)$ to store all active occupied MO pairs (i, j) [SCHEME 3.7]. The nested loops can be now fused together [line 5] to guarantee load balance in the threading region.

SCHEME 3.7 Loop fusion

```
1  !$omp threadprivate(Ẽ(2))
2  !$OMP PARALLEL shared(E(2))
3  !$omp do nowait schedule(DYNAMIC)
4  //fusing loops using the buffer array, tri
5  do k = 1, nocc × (nocc − 1)/2
6        i = tri(k, 1)
7        j = tri(k, 2)
8        Ẽ(2) ← 2.0 × corr(i, j, B_DV^i, B_DV^j)
9  enddo k
10 !$omp do nowait schedule(DYNAMIC)
11 do i = Lstart, Lend
12       Ẽ(2) ← 1.0 × corr(i, j, B_DV^i, B_DV^j)
13 enddo i
14 !$omp atomic
15       E(2) ← Ẽ(2)
16 !$OMP END PARALLEL
```

3.4 Computational models

The performance and scaling of the GDDI/OpenMP FMO/RI-MP2 implementation relative to the pure GDDI code is examined using water clusters of 139 (w139), 1120 (w1120) and 2165 (w2165) molecules shown in FIGURE 3.2. For comparison and simplicity, the RI auxiliary basis is taken to be cc-pVDZ-RI while the AO basis sets used are 3-21G, 6-31G(d) and cc-pVDZ. The calculations were done using the DOE 64-core KNL 7230 cluster.[26]

The computational model is denoted CODE/**N**//MOL/**n**/BASIS, where CODE=OMP or GDDI for GDDI/OpenMP or pure GDDI[14,19] FMO/RI-MP2 implementation, respectively. **N** is the number of KNL nodes used; MOL is a water cluster partitioned into **n** fragments; BASIS is the AO basis. For instance, OMP/**512**//w2165/**217**/cc-pVDZ is the GDDI/OpenMP FMO/RI-MP2 energy calculation for a water cluster of 2165 molecules (w2165) partitioned into 217 fragments using the cc-pVDZ/cc-pVDZ-RI basis set and 512 compute nodes (~33,000 cpus).

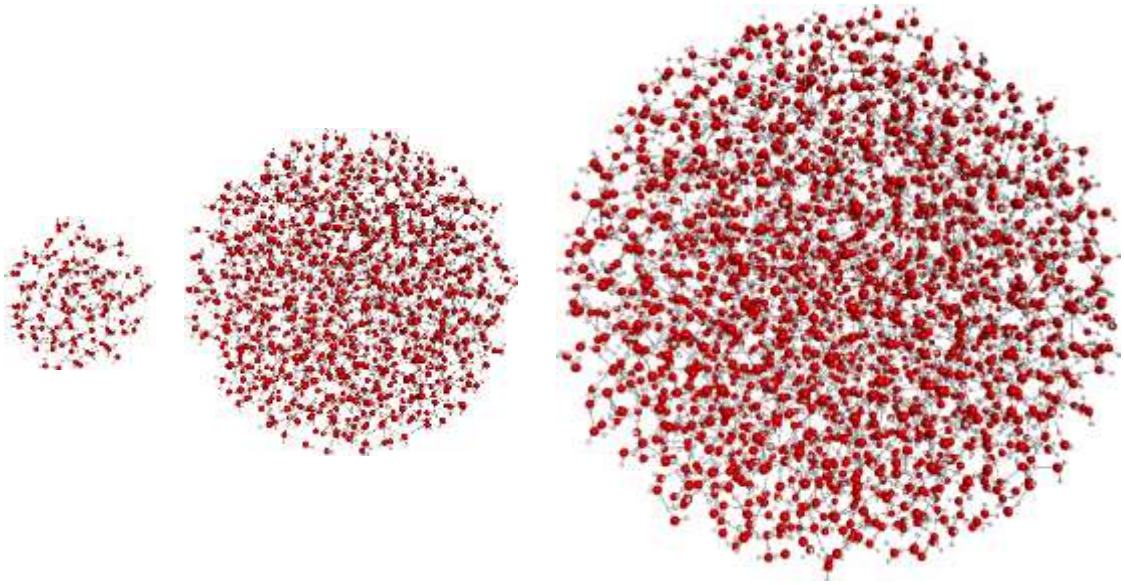


FIGURE 3.2 Clusters of 139 (w139), 1120 (w1120) and 2165 (w2165) water molecules.

3.5 Results and discussion

3.5.1 Ranks Per Node (***RPN***)

The wall time of OMP/**8**//w139/**14**/BASIS (BASIS=3-21G, 6-31G(d) and cc-pVDZ) calculations for various RPN are recorded in TABLE 3.1, in which each rank spawned a team of TPR threads so that all physical cores are used (i.e., $RPN \times TPR = 64$). The corresponding GDDI/**8**//w139/**14**/BASIS calculations with $RPN = 64$ are also carried out for comparison. The best OMP code performance was achieved when $RPN = 1$ as expected, whereas increasing RPN introduces more communication overhead and splitting on-node shared data structures increases the wall clock time. The optimal OMP calculations $(RPN = 1)$ are also about an order of magnitude faster than the pure GDDI calculations.

TABLE 3.1 The variation of wall clock time (s) with respect to RPN and TPR

$\boldsymbol{RPN}$	$\boldsymbol{TPR}$	OMP/**8**//w139/**14**/BASIS		
		3-21G	**6-31G(d)**	**cc-pVDZ**
1	**64**	14.3	25.7	44.9
2	**32**	38.1	75.9	130.7
4	**16**	39.7	77.2	135.6
8	**8**	41.8	81.4	140.5
GDDI/**8**//w139/**14**/BASIS		211.9	330.5	518.2

3.5.2 Threads Per Rank ($\boldsymbol{TPR}$) and relative performance.

The node resource usage is examined using the OMP and GDDI/**8**//w139/**14**/BASIS (BASIS=3-21G, 6-31G(d) and/or cc-pVDZ) calculations. For OMP calculations, RPN is set to 1 and TPR varies from $1-64$. In the corresponding pure GDDI calculations, the RPN was varied from $1-64$. The results for BASIS=6-31G(d) are depicted in FIGURE 3.3; Similar results for other basis sets are shown in the Supporting Information (SI). For the same computational resources used, [i.e., $RPN \times TPR \times$ N ($\#\ nodes$) in the OMP code equals $RPN \times$ N ($\#\ nodes$) in the GDDI code], the OMP wall time is found to be much smaller than that for GDDI. The OMP speedup relative to GDDI increases when the number of cores used is increased; when $TPR^{OMP} = RPN^{GDDI} = 64$ the relative speedup achieves a factor of $10\times$ [FIGURE 3.3b]. The OMP calculation wall time keeps decreasing as the TPR increases, whereas the GDDI wall time is soon saturated when the RPN increases to ½ the number of cores on a node. Node resources are, therefore, used more efficiently in the OMP implementation.

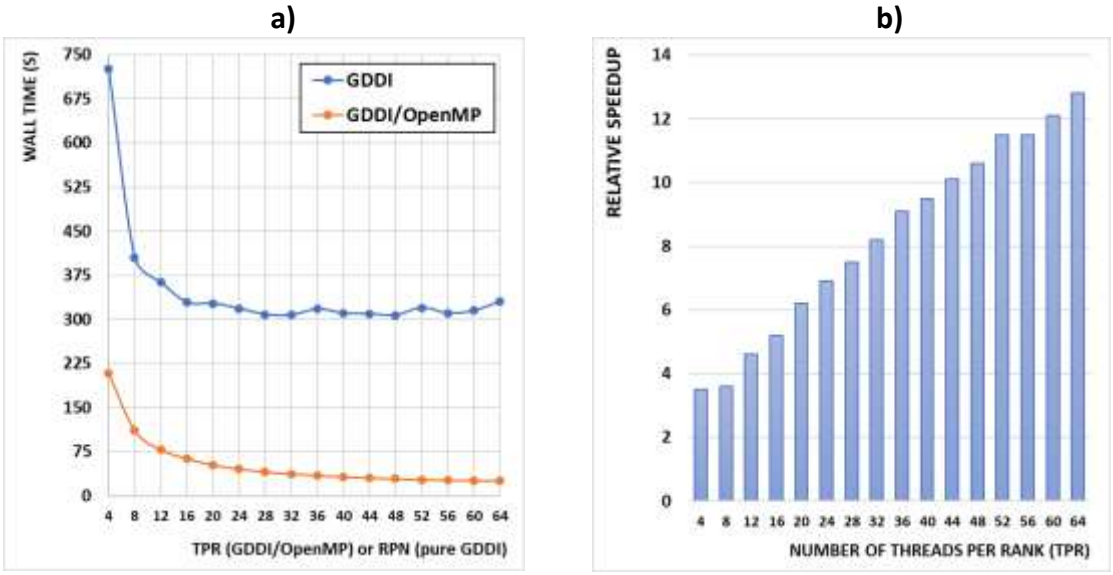


FIGURE 3.3 a) wall time and b) OMP speedup relative to the GDDI calculations.

The relative performance is also examined for larger calculations using OMP/**512**//MOL/**n**/BASIS, in which MOL are w1120 ($\boldsymbol{n = 112}$), w2165 ($\boldsymbol{n = 217}$); BASIS=6-31G(d) and cc-pVDZ. The results summarized in TABLE 3.2 show that the OMP code remains an order of magnitude (10x) faster than pure GDDI calculations.

TABLE 3.2 Wall time (s) for CODE/**512**//MOL_**n**/BASIS calculations. For CODE=OMP, RPN=1 and TPR=64; for CODE=GDDI, RPN=64.

MOL/**n**	w1120/**112**		w2165/**217**	
CODE	OMP	GDDI	OMP	GDDI
6-31G(d)	27.8	348.3	101.8	1298.5
cc-pVDZ	47.8	546.7	173.5	2032.7

3.5.3 Node Scaling.

Since the FMO method is well known to exhibit linear scaling,[27,28] and since the OMP FMO/RI-MP2 calculations for small to large systems [w139-w2165] using small to large numbers of cpus [512-33,000] exhibit an order of magnitude speedup relative to GDDI, the OMP FMO/RI-MP2

implementation is expected to scale well. FIGURE 3.4 depicts the variation of wall time and wall time relative to a 256-node job for OMP/**N**//w2165/**217**/6-31G(d) calculations [**N**=256-768].

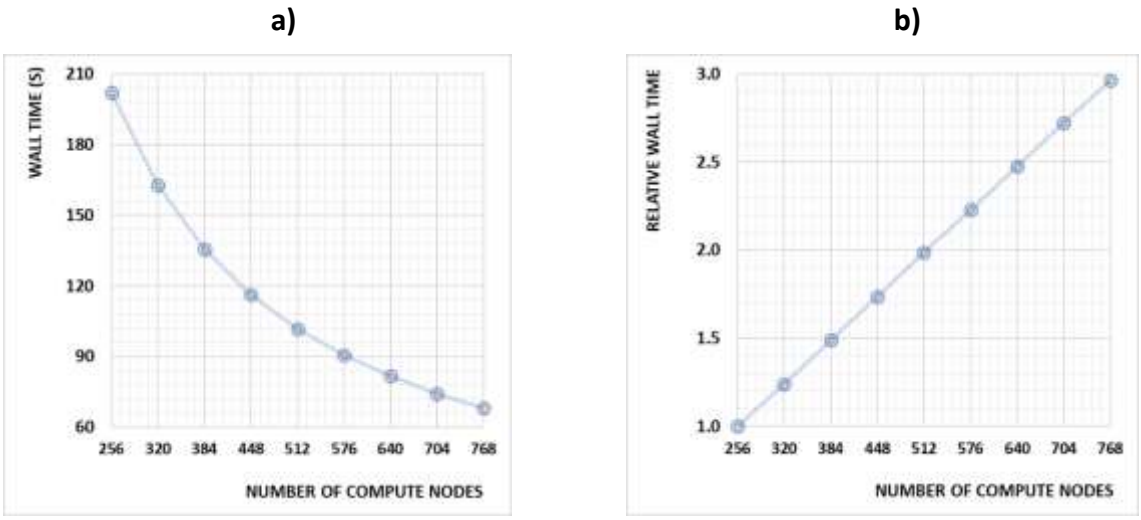


FIGURE 3.4 a) wall time and b) wall time relative to a 256-node job for OMP/ **N**//w2165/**217**/6-31G(d) calculations. **N** is the number of compute nodes.

3.6 Concluding remarks

The hybrid distributed/shared memory parallel model (GDDI/OpenMP) can make efficient and effective use of multiprocessor resources (on-node memory, physical cores) to speed up FMO/RI-MP2 calculations. Testing on medium size water clusters has demonstrated that the wall times of pure GDDI FMO/RI-MP2 calculations remain unchanged when using more than 1/2 of the KNL node resources (e.g., above 32 cores over 64 physical cores per node). In contrast, the wall time for GDDI/OpenMP FMO/RI-MP2 calculations monotonically decreases up to all KNL physical cores used. This results in a speedup of GDDI/OpenMP FMO/RI-MP2 relative to pure GDDI by a factor of 10x for calculations on medium to large molecular systems (e.g., of 417-11,259 atoms). In addition, the hybrid parallel model can also fully preserve the linear scaling

characteristics of the FMO framework. Calculations for water clusters that contain 2,165 molecules (11,259 atoms) exhibits linear scaling when the number of compute nodes is varied from 256-768.

Acknowledgements. This work was supported an Exascale Computing Project (ECP) grant from the Department of Energy, administered by the Ames Laboratory, Iowa State University. The authors gratefully acknowledge the Argonne Leadership Computing Facility (ALCF) at the Argonne National Laboratory (ANL) and the National Energy Research Scientific Computing Center (NERSC) for providing CPU time and technical support.

References

(1) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. Pair Interaction Molecular Orbital Method: An Approximate Computational Method for Molecular Interactions. *Chem. Phys. Lett.* **1999**, *312*, 319–324.

(2) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment Molecular Orbital Method: An Approximate Computational Method for Large Molecules. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(3) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment Molecular Orbital Method: Application to Polypeptides. *Chem. Phys. Lett.* **2000**, *318*, 614–618.

(4) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment Molecular Orbital Method: Use of Approximate Electrostatic Potential. *Chem. Phys. Lett.* **2002**, *351*, 475–480.

(5) Nagata, T.; Fedorov, D. G.; Kitaura, K. Mathematical Formulation of the Fragment Molecular Orbital Method BT - Linear-Scaling Techniques in Computational Chemistry and Physics: Methods and Applications; Zalesny, R., Papadopoulos, M. G., Mezey, P. G., Leszczynski, J., Eds.; Springer Netherlands: Dordrecht, 2011; pp 17–64.

(6) Mochizuki, Y.; Koikegami, S.; Nakano, T.; Amari, S.; Kitaura, K. Large Scale MP2 Calculations with Fragment Molecular Orbital Scheme. *Chem. Phys. Lett.* **2004**, *396*, 473–479.

(7) Mochizuki, Y.; Nakano, T.; Koikegami, S.; Tanimori, S.; Abe, Y.; Nagashima, U.; Kitaura, K. A Parallelized Integral-Direct Second-Order Møller–Plesset Perturbation Theory Method with a Fragment Molecular Orbital Scheme. *Theor. Chem. Acc.* **2004**, *112*, 442–452.

(8) Fedorov, D. G.; Ishimura, K.; Ishida, T.; Kitaura, K.; Pulay, P.; Nagase, S. Accuracy of the Three-Body Fragment Molecular Orbital Method Applied to Møller–Plesset Perturbation Theory. *J. Comput. Chem.* **2007**, *28*, 1476–1484.

(9) Mochizuki, Y.; Yamashita, K.; Murase, T.; Nakano, T.; Fukuzawa, K.; Takematsu, K.; Watanabe, H.; Tanaka, S. Large Scale FMO-MP2 Calculations on a Massively Parallel-Vector Computer. *Chem. Phys. Lett.* **2008**, *457*, 396–403.

(10) Fedorov, D. G.; Kitaura, K. Second Order Møller-Plesset Perturbation Theory Based upon the Fragment Molecular Orbital Method. *J. Chem. Phys.* **2004**, *121*, 2483–2490.

(11) Nagata, T.; Fedorov, D. G.; Ishimura, K.; Kitaura, K. Analytic Energy Gradient for Second-Order Møller-Plesset Perturbation Theory Based on the Fragment Molecular Orbital Method. *J. Chem. Phys.* **2011**, *135*, 44110.

(12) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(13) Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. On Some Approximations in Applications of Xα Theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(14) Ishikawa, T.; Kuwata, K. Fragment Molecular Orbital Calculation Using the RI-MP2 Method. *Chem. Phys. Lett.* **2009**, *474*, 195–198.

(15) Gordon, M. S.; Schmidt, M. W. Chapter 41 - Advances in Electronic Structure Theory: GAMESS a Decade Later.; Frenking, G., Kim, K. S., Scuseria, G. E. B. T.-T. and A. of C. C., Eds.; Elsevier: Amsterdam, 2005; pp 1167–1189.

(16) The General Atomic and Molecular Electronic Structure System (GAMESS 2018-R3). https://www.msg.chem.iastate.edu/gamess/.

(17) Ishikawa, T.; Kuwata, K. RI-MP2 Gradient Calculation of Large Molecules Using the Fragment Molecular Orbital Method. *J. Phys. Chem. Lett.* **2012**, *3*, 375–379.

(18) Ishikawa, T.; Ishikura, T.; Kuwata, K. Theoretical Study of the Prion Protein Based on the Fragment Molecular Orbital Method. *J. Comput. Chem.* **2009**, *30*, 2594–2601.

(19) Fedorov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. A New Hierarchical Parallelization Scheme: Generalized Distributed Data Interface (GDDI), and an Application to the Fragment Molecular Orbital Method (FMO). *J. Comput. Chem.* **2004**, *25*, 872–880.

(20) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral Approximations for LCAO-SCF Calculations. *Chem. Phys. Lett.* **1993**, *213*, 514–518.

(21) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: Optimized Auxiliary Basis Sets and Demonstration of Efficiency. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

(22) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. The Distributed Data Interface in GAMESS. *Comput. Phys. Commun.* **2000**, *128*, 190–200.

(23) OpenMP. https://www.openmp.org/.

(24) Pas, R. van der; Stotzer, E.; Terboven, C. *Using OpenMP - the next Step : Affinity,*

Accelerators, Tasking, and SIMD; 2017.

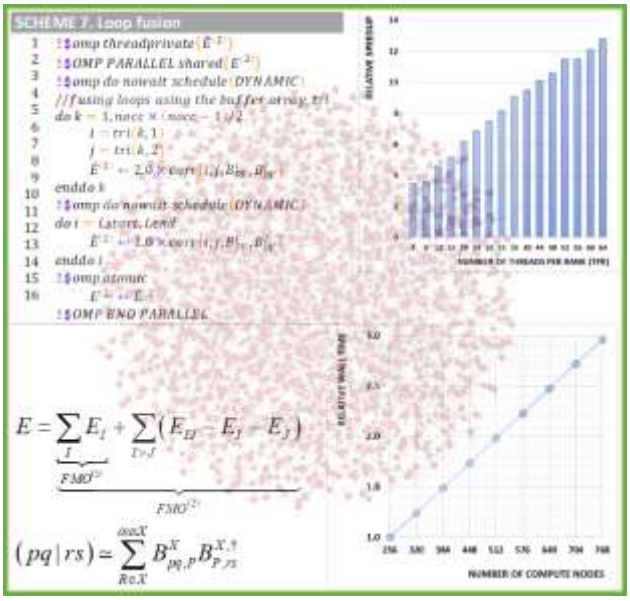
(25) Clauss, P.; Altintas, E.; Kuhn, M. Automatic Collapsing of Non-Rectangular Loops. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*; 2017; pp 778–787.

(26) Theta Computer at ALFC. https://www.alcf.anl.gov/theta.

(27) Fedorov, D. G. The Fragment Molecular Orbital Method: Theoretical Development, Implementation in GAMESS, and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1322.

(28) Pruitt, S. R.; Nakata, H.; Nagata, T.; Mayes, M.; Alexeev, Y.; Fletcher, G.; Fedorov, D. G.; Kitaura, K.; Gordon, M. S. Importance of Three-Body Interactions in Molecular Dynamics Simulations of Water Demonstrated with the Fragment Molecular Orbital Method. *J. Chem. Theory Comput.* **2016**, *12*, 1423–1435.

Graphical abstract



CHAPTER 4. A MULTI-LEVEL PARALLEL IMPLEMENTATION OF FMO/RI-MP2 ANALYTIC GRADIENT

A paper on preparation for publication at a later date.

Buu Q. Pham and Mark S. Gordon

Abstract

A hybrid multi-level parallel approach, based on a combination of the general distributed data interface (GDDI) combined with shared memory OpenMP API, is applied to the analytic gradient of the resolution-of-the-identity (RI) second-order Moller-Plesset perturbation theory (MP2) within the fragment molecular orbital (FMO) method. The FMO/RI-MP2 analytic gradient is derived and implemented in the GAMESS electronic structure suite of programs. The MP2 gradient in the FMO framework contains *three* parts; i.e., the internal component, the electrostatic potential (ESP) component and the response terms. The MP2 density matrices and internal fragment Lagrangian are generated in the internal gradient driver, which are shared with the ESP and response contributions to the gradient. In this paper, the RI approximation is applied to the MP2 density matrices, the MP2 amplitude, the 4-index 2-electron repulsion integral derivative coupled with the MP2 amplitude, and the internal fragment Lagrangian. The FMO/RI-MP2 analytic gradient is validated against the numerical gradient, and the method is demonstrated with molecular dynamics (MD) simulations using an NVE ensemble for water clusters. The accuracy and performance relative to full FMO/MP2 are also examined using water clusters of varying sizes. The maximum FMO/RI-MP2 analytic and numerical gradient difference is in the range of $10^{-5} - 10^{-6}$ Hartree/Bohr. The log-log plot of RMSD(E) against the time step

in NVE ensemble is close to the ideal value of 2.0. The FMO/RI-MP2 gradient introduces a small error relative to the full FMO/MP2 gradient (e.g., maximum gradient error of 10^{-5} Hartree/Bohr) and significantly speeds up the calculations by a factor of 3.9-8.0x.

4.1 Introduction

Ab initio calculations for large molecular systems are computationally demanding. For instance, the formal computational scaling of some popular wave function based methods vary from $\mathcal{O}(N^4)$ for Hartree-Fock (HF), $\mathcal{O}(N^5)$ for second-order Moller-Plesset perturbation theory (MP2), up to $\mathcal{O}(N^7)$ or more for coupled cluster (CC) methods. Here, N is the size of the molecular system, which can be the number of atoms or the number of basis functions. The increase in molecular system size in *ab initio* calculations introduces not only a steep rise in computational cost and memory demand, but also increasing difficulty in developing efficient parallel implementations that can take advantage of very large computing resources. In fragmentation methods, a system is typically partitioned into small fragments, which can be treated relatively independently. This can reduce computational cost and memory demands and facilitate parallel implementation, yielding good scaling in large calculations.

The fragment molecular orbital (FMO) method[1–5] has been demonstrated to be a linear scaling many-body expansion method that maintains an accuracy that is close to the original underlying method. Many *ab initio* electronic structure methods (e.g., HF, MP2, CC) and density functional theory methods have been integrated into the FMO framework. In the GAMESS electronic structure package,[6] FMO combined with most electronic structure methods can achieve node linear scaling due to the support of the general distributed data interface (GDDI)[7]

that efficiently distributes computing resources (e.g., computing nodes) among fragments. While the FMO method can achieve good scaling, the computational cost remains large for big molecular systems, particularly when using a method that accounts for dynamic correlation effects, such as MP2.

Several strategic approaches have been devised to facilitate large correlated FMO calculations. One approach is to exploit the local nature of many large molecular systems to reduce the computational cost by, for example, employing the electrostatic potential (ESP) to represent distant inter-fragment interactions rather than full *ab initio* calculations.[4,8] In the same vein, the recently developed effective fragment molecular orbital (EFMO) method[9,10] treats distant dimers with the effective potential (EFP) model. [11] A second approach is to apply approximations to reduce the computational cost of the fragment *ab initio* calculation itself. These include the use of the Cholesky decomposition (CD) method[12–15] or the resolution-of-the-identity (RI) approximation[16–18] to reduce the computational cost and memory storage of 4-index 2-electron repulsion integrals (4-2ERIs) in the molecular orbital (MO) basis.[19–21] Finally, to make use of new multicore processor generations (e.g., Haswell or Knights Landing processors), it is important to replace the distributed memory model (e.g., plain GDDI) by more efficient parallel models (e.g., the hybrid distributed/shared memory model GDDI/OpenMP).[22,23] In this work, the RI approximation and the hybrid GDDI/OpenMP are used to speed up the FMO/MP2 analytic gradient.

The (analytic) gradient is necessary for probing potential energy surfaces (e.g., searching for stationary structures such as minima and saddle points), calculating forces in molecular dynamics (MD) simulations, and as a first step in predicting vibrational frequencies. In comparison with the

gradient of pure *ab initio* methods, the gradient of the *ab initio* methods integrated into the FMO framework have two distinct differences. In FMO methods, each (monomer) fragment *ab initio* calculation is embedded in the electrostatic potential due to the nuclei and electron densities of all other fragments (monomers). The monomer calculations are repeated until the charge distributions of all monomers are self-consistent, a procedure called the self-consistent charge (SCC) method. Following the convergence of the monomer charges, the exchange, charge transfer and other interactions among fragments are included by many-body corrections. Two- and/or three-body corrections require *ab initio* calculations for fragment pairs (dimers) and/or triples (trimers). The dimer and trimer calculations are embedded in the *fixed* ESP of the leaving monomers and are iterated until dimer or trimer charge distribution is (separately) self-consistent. The charge distribution of dimer (trimer) and the corresponding pair (triple-pair) of monomers are generally different. This subsequently introduces additional response terms in FMO/RHF gradient. For the FMO/RHF level of theory, the response terms can be manipulated and transformed into a collective monomer response term, which is then transformed into a Z-vector formalism[24] to reduce the number of equations that needs to be solved. The FMO Z-vector is obtained from the special self-consistent Z-vector (SCZV) solver first derived and implemented in GAMESS by Nagata et *al*.[25,26]

In non-variational electron correlation methods such as MP2, the fragment gradient also introduces response terms. Monomer MP2 response terms are *directly* added to the FMO/RHF response terms and solved by the SCZV solver without additional cost. For dimers and trimers, the MP2 response terms first must be transformed to monomer response terms using coupled-perturbed Hartree-Fock (CPHF) equations before entering the SCZV solver.[25,26] Such a

transformation requires solving for the dimer (and trimer) Z-vectors.[24] This paper briefly summarizes the FMO method, the CPHF equations, and the FMO/MP2 gradient before introducing the RI approximation to the appropriate terms, followed by a brief introduction of the hybrid GDDI/OpenMP parallel implementation. Benchmark calculations are carried out using water clusters and are compared with the regular GDDI FMO/MP2 gradient reference.

4.2 FMO2/RHF analytic gradient

4.2.1 Some general notation

In this paper, unless otherwise noted, the following notations are applied to the equations.

i) The entire MO space of a fragment X is denoted as $allX$. This can be split into the occupied $occX$ and virtual $virtX$ MO subspaces. The occupied MO subspace can be further divided into the core $corX$, and active occupied MOs $actX$. For the RI approximation, $auxX$ stands for the auxiliary basis of the fragment X. The atomic orbital basis of a fragment X is denoted as aoX.

ii) In equations, for brevity, the index p can be used in place of an MO $\varphi_p(r)$. The indices p, q, r, s represent general MOs in the $allX$ subspace; indices k, l, k', l' represent MOs in the *occX* subspace; I, J are in the $corX$ subspace; i, j are in the $actX$ subspace; and a, b, c, d represent the virtual MOs $virtX$.

iii) The Dirac notation is used for integrals. The 1-electron integrals include the kinetic energy integral defined in Eq. (1), the electron-nuclear attraction in Eq. (2), and the overlap integral in Eq. (3), in which r and r_k are electron coordinates; Z_α is the nuclear charge of nucleus α. The 2-electron integrals include the 4-index 2-electron repulsion

integral (4-2ERI) among MOs defined in Eq. (4), the 3-2ERI among two MOs $\{\varphi_p\}$ and one auxiliary basis function $\{\alpha_P\}$ in the RI approximation in Eq. (5), and 2-2ERI between two auxiliary basis functions in Eq. (6).

$$\left(p\left| -\frac{1}{2}\nabla^2(r_k)\right| q\right) = \int dr_k \varphi_p^*(r_k)\left(-\frac{1}{2}\nabla^2(r_k)\right)\varphi_q(r_k) \tag{1}$$

$$\left(p\left| -\sum_{\alpha}^{nuc}\frac{Z_\alpha}{R_{\alpha k}}\right| q\right) = \int dr_k \varphi_p^*(r_k)\left(-\sum_{\alpha}^{nuc}\frac{Z_\alpha}{R_{\alpha k}}\right)\varphi_q(r_k) \tag{2}$$

$$S_{pq} = \int dr \varphi_p^*(r)\varphi_q(r) \tag{3}$$

$$(pq|rs) = \iint dr_1 dr_2 \varphi_p^*(r_1)\varphi_q(r_1) r_{12}^{-1}\varphi_r^*(r_2)\varphi_s(r_2) \tag{4}$$

$$(pq|P) = \iint dr_1 dr_2 \varphi_p^*(r_1)\varphi_q(r_1) r_{12}^{-1}\alpha_P(r_2) \tag{5}$$

$$(P|Q) = \iint dr_1 dr_2 \alpha_P(r_1) r_{12}^{-1}\alpha_Q(r_2) \tag{6}$$

iv) In the FMO notation, the *tilde* stands for an internal fragment term (e.g., the internal Fock matrix element of the fragment X is $\tilde{F}_{pq}^X$), and the bar denotes an ESP-related term (e.g., the ESP-related Fock matrix element $\bar{F}_{pq}^X$).

v) The derivative of a matrix element (e.g., Fock matrix element F_{pq}^X) with respect to the nuclear displacement ζ is denoted by the superscript ζ (e.g., $F_{pq}^{X,\zeta}$); the derivative of a matrix element with respect to ζ in the AO basis transformed back to the MO basis is denoted by the superscript in the parentheses (ζ).

$$\frac{\partial}{\partial\zeta}F_{pq}^X \equiv F_{pq}^{X,\zeta} \tag{7}$$

$$\sum_{\mu\nu}^{aoX} C_{\mu p}^X \left(\frac{\partial}{\partial\zeta}F_{\mu\nu}^X\right) C_{\nu q}^X \equiv F_{pq}^{X,(\zeta)} \tag{8}$$

4.2.2 FMO2/RHF analytic gradient

The FMO2 energy can be formulated in terms of the monomer E_I and dimer E_{IJ} energy

$$E = \sum_I E_I + \sum_{I>J} \left(E_{IJ} - E_I - E_J \right) \tag{9}$$

The RHF energy of a general (monomer or dimer) fragment X is given by

$$E_X = \sum_{i \in X}^{occX} \left(\widetilde{H}_{ii}^X + \overline{F}_{ii}^X + F_{ii}^X \right) \tag{10}$$

The Fock matrix element F_{pq}^X contains the internal $\tilde{F}_{pq}^X$ and ESP $\bar{F}_{pq}^X$ components.

$$F_{pq}^X = \widetilde{F}_{pq}^X + \overline{F}_{pq}^X \tag{11}$$

Similar to an isolated fragment, the internal Fock matrix element $\tilde{F}_{pq}^X$ includes the 1-electron integral $\tilde{H}_{pq}^X$, the 2-electron internal Coulomb $\tilde{J}_{pq}^X$ and exchange $\tilde{K}_{pq}^X$ terms.

$$\widetilde{F}_{pq}^X = \widetilde{H}_{pq}^X + \widetilde{J}_{pq}^X - \widetilde{K}_{pq}^X \tag{12}$$

$$\widetilde{H}_{pq}^X = \left(p \left| -\frac{1}{2} \nabla^2 \left(r_k \right) - \sum_{\alpha \in X}^{nucX} \frac{Z_\alpha}{R_{\alpha k}} \right| q \right) \tag{13}$$

$$\widetilde{J}_{pq}^X - \widetilde{K}_{pq}^X = \sum_{k \in X}^{occX} \left[2\left(pq \,|\, kk \right) - \left(pk \,|\, kq \right) \right] \tag{14}$$

In Eqs. (53) and (54), $nucX$ and $occX$ are nuclei and occupied MOs in the fragment X, respectively. The $(pq|rs)$ are 4-2ERIs over the MO basis functions. The ESP Fock matrix element $\bar{F}_{pq}^X$ contains the Coulomb interactions between electrons in fragment X with nuclei $\bar{u}_{pq}^X$ and the electron density $\bar{J}_{pq}^X$ of all other fragments

$$\bar{F}_{pq}^{X} = \bar{u}_{pq}^{X} + \bar{J}_{pq}^{X} \tag{15}$$

$$\bar{u}_{pq}^{X} = \left(p \left| -\sum_{K \neq X} \sum_{\alpha \in K}^{nucK} \frac{Z_\alpha}{R_{\alpha k}} \right| q \right) \tag{16}$$

$$\bar{J}_{pq}^{X} = 2 \sum_{K \neq X} \sum_{k \in K}^{occK} (pq|kk) \tag{17}$$

In Eqs. (58) and (59), $nucK$ and $occK$ are nuclei and occupied MOs in fragment $K \neq X$, respectively. In addition to terms defined in Eqs. (52)-(59), additional terms must be added to the internal Fock matrix elements when the fragmentation breaks covalent bonds, which can be treated by either the hybrid orbital projection (HOP) method,[3,27] or by the adaptive frozen orbital (AFO) method.[28–30] Since only the FMO gradient with the HOP scheme has been completed to date, this paper only considers the HOP contribution to the Fock matrix. This is accomplished by P_{pq}^{X} defined in Eq. (55) through the HOP operator $\hat{P}^{X}$ in Eq. (56), in which θ_k is a hybrid orbital, and $B_k = 10^{6-8}$ is called the universal constant.[3,27]

$$P_{pq}^{X} = \left(p \left| \hat{P}^{X} \right| q \right) \tag{18}$$

$$\hat{P}^{X} = \sum_{k \in X} B_k \left| \theta_k \right\rangle \left\langle \theta_k \right| \tag{19}$$

Taking the derivative of the energy of fragment X with respect to a nuclear displacement ζ yields the gradient consisting of the internal gradient term $\tilde{E}_X^{\zeta}$, the ESP gradient term $\bar{E}_X^{\zeta}$, and the response term U_X^{ζ}.

$$\frac{\partial E^{X}}{\partial \zeta} = \tilde{E}_X^{\zeta} + \bar{E}_X^{\zeta} + U_X^{\zeta} \tag{20}$$

$$\tilde{E}_X^{\zeta} = \sum_{k\in X}^{occX}\left[\tilde{H}_{kk}^{X,(\zeta)} + \tilde{F}_{kk}^{X,(\zeta)}\right] - 2\sum_{k\in X}^{occX} S_{kk}^{X,(\zeta)}\varepsilon_k^X \tag{21}$$

$$\bar{E}_X^{z} = 2\sum_{k\in X}^{occX}\bar{F}_{kk}^{X,(z)} - 4\sum_{k\in X}^{occX}\sum_{K\neq X}\sum_{kl\in K}^{occK} S_{kl}^{K,(z)}\left(k'k'|kl\right) \tag{22}$$

$$U_X^z = 8\sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,z}\sum_{k'\in X}^{occX}\left(k'k'|ck\right) \tag{23}$$

The *total* FMO2/RHF energy gradient also includes the separate internal gradient term $\Delta\tilde{E}^{\zeta}$, the ESP gradient term $\Delta\bar{E}^{\zeta}$, and the response gradient term ΔU^{ζ}

$$\frac{\partial E}{\partial\zeta} = \Delta\tilde{E}^{\zeta} + \Delta\bar{E}^{\zeta} + \Delta U^{\zeta} \tag{24}$$

$$\Delta\tilde{E}^{\zeta} = \sum_I \tilde{E}_I^{\zeta} + \sum_{I>J}\left(\tilde{E}_{IJ}^{\zeta} - \tilde{E}_I^{\zeta} - \tilde{E}_J^{\zeta}\right) \tag{25}$$

$$\Delta\bar{E}^{\zeta} = \sum_I \bar{E}_I^{\zeta} + \sum_{I>J}\left(\bar{E}_{IJ}^{\zeta} - \bar{E}_I^{\zeta} - \bar{E}_J^{\zeta}\right) \tag{26}$$

$$\Delta U^{\zeta} = \sum_I U_I^{\zeta} + \sum_{I>J}\left(U_{IJ}^{\zeta} - U_I^{\zeta} - U_J^{\zeta}\right) \tag{27}$$

By plugging Eq. (23) into (27), and after some algebra, the response term ΔU^{ζ} becomes the collective monomer response term shown in Eq. (28) with the Lagrangian $\boldsymbol{\mathcal{L}}^K$ and the density D defined in Eqs. (29) and (30), respectively. By examining the Lagrangian in Eq. (29), it can be seen that the response term in FMO2/RHF is due to the difference in charge distribution of dimers and the corresponding pair of monomers. The response term is then transformed into the Z-vector, which is obtained from the SCZV solver.[25] These steps are briefly discussed in the next sections.

$$\Delta U^{\zeta} = \sum_K\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta}\boldsymbol{\mathcal{L}}_{ck}^K \tag{28}$$

$$\mathcal{L}_{ck}^{K} = 4\sum_{IJ\neq K}\sum_{\mu\nu\in IJ}^{aoIJ}\left(\mu\nu\,|\,ck\right)\left[D_{\mu\nu}^{IJ} - \left(D_{\mu\nu}^{I}\oplus D_{\mu\nu}^{J}\right)\right] \tag{29}$$

$$D_{\mu\nu}^{X} = 2\sum_{k\in X}^{occX} C_{\mu k}^{X} C_{\nu k}^{X} \tag{30}$$

4.2.3 CPHF and CPHF-related equations

With the development of the Z-vector technique, the (FMO) CPHF and CPHF-related equations are merely used to transform MO response blocks into the Z-vector; they are additionally used as intermediate equations during the derivation of correlation energy derivatives. These equations can be obtained by differentiating the Fock matrix element in Eq. (51) that yields zero for off-diagonal element $(p \neq q)$, or the derivative of MO energy $\varepsilon_p^{X,\zeta}$ for diagonal elements.

$$\frac{\partial}{\partial\zeta}F_{pq}^{X} = B_{pq}^{X,(\zeta)} + U_{pq}^{X,\zeta}\left(\varepsilon_p^{X} - \varepsilon_q^{X}\right) + \sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta} A_{pq,ck}^{X,X} + \sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta} A_{pq,ck}^{X,K} = \delta_{pq}\varepsilon_p^{X,\zeta} \tag{31}$$

The upper limits on the summations $virtX$ and $virtK$ are the virtual MO spaces of fragments X and K, respectively. The term $B_{pq}^{X,(\zeta)}$ consists of the derivative of the Fock and overlap matrix elements in the AO basis transformed back to the MO basis, and the matrix A consists of sequences of two-electron integrals:

$$B_{pq}^{X,(\zeta)} = F_{pq}^{X,(\zeta)} - S_{pq}^{X,(\zeta)}\varepsilon_q^{X} - \sum_{kl\in X}^{occX} S_{kl}^{X,(\zeta)}\,\tilde{A}_{pq,kl}^{X,X} - \frac{1}{2}\sum_{K\neq X}\sum_{kl\in X}^{occX} S_{kl}^{K,(\zeta)} A_{pq,kl}^{X,K} \tag{32}$$

$$A_{pq,rs}^{X,X} = 4\left(pq\,|\,rs\right) - \left(pr\,|\,sq\right) - \left(ps\,|\,rq\right);\ p,q,r,s\in X \tag{33}$$

$$\tilde{A}_{pq,kl}^{X,X} = 2\left(pq\,|\,kl\right) - \left(pk\,|\,lq\right);\ p,q,k,l\in X \tag{34}$$

$$A_{pq,rs}^{X,K} = 4(pq|rs);\ p,q \in X;\ r,s \in K \tag{35}$$

From Eq. (31), the MO response block $U_{pq}^{X,\zeta}$ can be written in a form, Eq. (36), that is useful for the derivation of the correlation energy gradient.

$$U_{pq}^{X,\zeta} = \frac{Q_{pq}^{X,(\zeta)}}{\varepsilon_q^X - \varepsilon_p^X} \tag{36}$$

$$Q_{pq}^{X,(\zeta)} = B_{pq}^{X,(\zeta)} + \sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta} A_{pq,ck}^{X,X} + \sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta} A_{pq,ck}^{X,K} \tag{37}$$

Eq. (36) is valid in practical use when the singularity $\left(\varepsilon_q^X - \varepsilon_p^X = 0\right)$ can be removed; or when the pairwise MO energy difference is not close to zero (e.g., the core-active occupied MO block). For diagonal Fock matrix elements, Eq. (31) gives the MO energy derivative $\varepsilon_p^{X,\zeta}$

$$\frac{\partial}{\partial\zeta} F_{pp}^X \equiv \varepsilon_p^{X,\zeta} = Q_{pp}^{X,(\zeta)} \tag{38}$$

For the virt-occ block of the MO response, Eq. (31) can be written in the form of Eq. (39) or in the matrix form in Eq. (40), in which δ is the Kronecker delta. These equations are useful when transforming the virt-occ block of the *dimer* MO response to the *monomer* MO response term in correlated FMO methods.

$$\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta} \overset{\approx}{A}{}_{dl,ck}^{X,X} = \tilde{B}{}_{dl}^{X,(\zeta)} \tag{39}$$

The quantities in Eq. (39) are defined as follows:

$$\overset{\approx}{A}{}^{X,X,\dagger} U^{X,\zeta} = \tilde{B}{}^{X,(\zeta)} \tag{40}$$

$$\overset{\approx}{A}{}_{dl,ck}^{X,X} = \delta_{cd}\delta_{kl}\left(\varepsilon_l^X - \varepsilon_d^X\right) - A_{dl,ck}^{X,X} \tag{41}$$

$$\tilde{B}_{dl}^{X,(\zeta)} = B_{dl}^{X,(\zeta)} + \sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta} A_{dl,ck}^{X,K} \tag{42}$$

Finally, for the virt-occ block of the MO response, Eq. (31) can also be transformed to the well-known CPHF equation shown Eq. (43), or in the matrix form in Eq. (44). The CPHF equation is usually used as a mean to transform the monomer MO response into the Z-vector, which is introduced in the next section.

$$\sum_{K}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,z} Q_{dl,ck}^{X,K} = B_{dl}^{X,(z)} \tag{43}$$

$$Q^{\dagger} U^{z} = B^{(z)} \tag{44}$$

$$\Theta_{dl,ck}^{X,K} = \delta_{XK}\tilde{\tilde{A}}_{dl,ck}^{X,X} + 4\left(\delta_{XK}-1\right) A_{dl,ck}^{X,K} \tag{45}$$

4.2.4 Z-vector method

Two MO response terms can arise when differentiating a general fragment X correlation energy in the correlated FMO/MP2 method. The first one is the *monomer* [if *X* is a monomer] or the *dimer* [if X is a dimer] internal response term $\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X}$; the second is the collective *monomer* MO response term $\sum_{K}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{K}$. If the fragment *X* is a monomer the first term can be merged into the second. However, if X is a *dimer*, the first term needs to be converted into the *monomer* response term using Eqs. (39)-(42) before it is merged into the second type of response term.

4.2.5 Dimer MO response term

The *dimer* MO response terms $\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X}$ can be converted into the *monomer* MO response terms using Eqs. (39)-(42) using the following procedure

$$\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X} \equiv \boldsymbol{\mathcal{L}}^{X,\dagger}U^{X,\zeta} = \boldsymbol{\mathcal{L}}^{X,\dagger}\left(\overset{\approx}{A}^{X,X,\dagger}\right)^{-1}\tilde{B}^{X,(\zeta)} = Z^{X,VO,\dagger}\tilde{B}^{X,(\zeta)} \equiv \sum_{d\in X}^{virtX}\sum_{l\in X}^{occX}\tilde{B}_{dl}^{X,(\zeta)}Z_{dl}^{X,VO} \quad (46)$$

The quantity $\tilde{B}^{X,\zeta}$ is defined in Eq. (42). The *dimer*-to-*monomer* MO response transformation needs the *dimer* Z-vector $Z^{X,VO}$, which is obtained by solving the regular *dimer* Z-vector equations[24] as shown in Eq. (47). The superscript VO in the Z-vector stands for the *virt-occ* block of the MO index; the dagger (†) indicates the transpose operation.

$$Z^{X,VO,\dagger} = \boldsymbol{\mathcal{L}}^{X,\dagger}\left(\overset{\approx}{A}^{X,X,\dagger}\right)^{-1} \Rightarrow \overset{\approx}{A}^{X,X} Z^{X,VO} = \boldsymbol{\mathcal{L}}^{X} \quad (47)$$

When the *dimer* Z-vector is available, the *dimer*-to-*monomer* transformation is given by Eq. (48). As shown in the later sections, in the FMO/MP2 gradient in GAMESS, the *dimer* Z-vectors are not passed directly to the monomer response term, but in the form of the MP2 density.

$$\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X} = \sum_{d\in X}^{virtX}\sum_{l\in X}^{occX} B_{dl}^{X,(\zeta)}Z_{dl}^{X,VO} + \sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta}\sum_{d\in X}^{virtX}\sum_{l\in X}^{occX} Z_{dl}^{X,VO}A_{dl,ck}^{X,K} \quad (48)$$

4.2.6 Collective monomer response term

The collective monomer response terms $\sum_K\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{K}$ can be transformed into the Z-vector terms using the following procedure (see Eqs. (43)-(45))

$$\sum_{K}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK} U_{ck}^{K,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{K} \equiv \mathbf{L}^{\dagger}U^{\zeta} = \boldsymbol{\mathcal{L}}^{\dagger}\left(\Theta^{\dagger}\right)^{-1}B^{(\zeta)} = Z^{\dagger}B^{(\zeta)} = \sum_{K}\sum_{d\in K}^{virtK}\sum_{l\in K}^{occK} B_{dl}^{X,(\zeta)}Z_{dl}^{K} \quad (49)$$

The collective monomer response term contribution to the FMO gradient in Eq. (49) is obtained after solving the Z-vector equation (Eq. (50)) using the SCZV solver, whose procedure described in detailed by Nagata et *al*.[25,26]

$$Z^{\dagger} = \mathcal{L}^{\dagger}\left(\Theta^{\dagger}\right)^{-1} \Rightarrow \Theta Z = \mathcal{L} \tag{50}$$

4.3 FMO2/MP2 analytic gradient

The dynamic correlation effect can be introduced into the FMO framework by adding the MP2 correlation energy $E_X^{(2)}$ into the fragment energy; e.g., $E_X \rightarrow E_X + E_X^{(2)}$. The MP2 correlation energy of the fragment X obtained by correlating the motion of the electrons in the active occupied MO space $(actX)$ using the virtual MO space $(virtX)$ can be formulated in terms of the 4-2ERIs in the MO basis $(pq|rs)$, the MP2 amplitude T_{ab}^{ij}, and the MO energy ε_p^X as follows.

$$E_X^{(2)} = \sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX}(ia\,|\,jb)T_{ab}^{ij} \tag{51}$$

$$T_{ab}^{ij} = 2t_{ab}^{ij} - t_{ab}^{ij,\dagger} \tag{52}$$

$$t_{ab}^{ij} = \frac{(ia\,|\,jb)}{D_{ab}^{ij}} \tag{53}$$

$$D_{ab}^{ij} = \varepsilon_i^X + \varepsilon_j^X - \varepsilon_a^X - \varepsilon_b^X \tag{54}$$

4.3.1 Fragment MP2 correlation energy gradient

Taking the derivative of the fragment MP2 correlation energy with respect to the nuclear displacement ζ (and after some algebra)[26,31] gives.

$$\begin{aligned}\frac{\partial}{\partial\zeta}E_X^{(2)} &= 2\sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX}\left(ia\,|\,jb\right)^{(\zeta)}T_{ab}^{ij} \\ &-2\sum_{i'i\in X}^{actX}S_{i'i}^{X,(\zeta)}L_{i'i}-4\sum_{k\in X}^{occX}\sum_{c\in X}^{virtX}S_{kc}^{X,(\zeta)}\widetilde{L}_{kc}-2\sum_{a'a\in X}^{virtX}S_{a'a}^{X,(\zeta)}\widetilde{L}_{a'a} \\ &+4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}\frac{Q_{Ii}^{X,(\zeta)}+Q_{iI}^{X,(\zeta)}}{2}P_{Ii}^{CA}-2\sum_{i'i\in X}^{actX}\frac{Q_{i'i}^{X,(\zeta)}+Q_{ii'}^{X,(\zeta)}}{2}P_{i'i}^{AA} \\ &+2\sum_{a'a\in X}^{virtX}\frac{Q_{a'a}^{X,(\zeta)}+Q_{aa'}^{X,(\zeta)}}{2}P_{a'a}^{VV}+4\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}U_{ck}^{X,\zeta}\left(4\delta_{kAct}L_{ck}-4\widetilde{L}_{kc}\right)\end{aligned} \tag{55}$$

In the 5th term, $corX$ stands for the core MOs in fragment X. In the density matrix notations (e.g., P^{CA}, P^{AA}, P^{VV}) the superscripts C, A, V stands for core,active occupied and virtual blocks of MOs, respectively. In the 8th term, $\delta_{kAct}=1$ when k is an active occupied MO index, otherwise it is *zero*. The density and energy-weighted density matrices in the MO basis are defined in Eqs. (56)-(60)

$$L_{mi}=\sum_{j\in X}^{actX}\sum_{ab\in X}^{virtX}\left(ma\,|\,jb\right)T_{ab}^{ij} \tag{56}$$

$$\widetilde{L}_{ma}=\sum_{ij\in X}^{actX}\sum_{a\in X}^{virtX}\left(im\,|\,jb\right)T_{ab}^{ij} \tag{57}$$

$$P_{Ii}^{CA}=\frac{L_{Ii}}{e_i^X-e_I^X} \tag{58}$$

$$P_{i'i}^{AA}=\sum_{j\in X}^{actX}\sum_{ab\in X}^{virtX}\frac{\left(i'a\,|\,jb\right)}{D_{ab}^{i'j}}T_{ab}^{ij} \tag{59}$$

$$P_{a'a}^{VV}=\sum_{ij\in X}^{actX}\sum_{b\in X}^{virtX}\frac{\left(ia'\,|\,jb\right)}{D_{a'b}^{ij}}T_{ab}^{ij} \tag{60}$$

Expanding $Q_{pq}^{X,(\zeta)}$ using Eq. (37), the fragment correlation energy gradient becomes

$$\begin{aligned}\frac{\partial}{\partial\zeta}E_X^{(2)} &= 2\sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX}\left(ia\,|\,jb\right)^{(\zeta)}T_{ab}^{ij} \\ &-2\sum_{i'i\in X}^{actX}S_{i'i}^{X,(\zeta)}L_{i'i}-4\sum_{k\in X}^{occX}\sum_{c\in X}^{virtX}S_{kc}^{X,(\zeta)}\widetilde{L}_{kc}-2\sum_{a'a\in X}^{virtX}S_{a'a}^{X,(\zeta)}\widetilde{L}_{a'a} \\ &+4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}\frac{B_{Ii}^{X,(\zeta)}+B_{iI}^{X,(\zeta)}}{2}P_{Ii}^{CA}-2\sum_{i'i\in X}^{actX}\frac{B_{i'i}^{X,(\zeta)}+B_{ii'}^{X,(\zeta)}}{2}P_{i'i}^{AA} \\ &+2\sum_{a'a\in X}^{virtX}\frac{B_{a'a}^{X,(\zeta)}+B_{aa'}^{X,(\zeta)}}{2}P_{a'a}^{VV}+\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}U_{ck}^{X,\zeta}\tilde{L}_{ck}^{X,X}+\sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK}U_{ck}^{K,\zeta}\bar{L}_{ck}^{X,K}\end{aligned} \tag{61}$$

The internal fragment Lagrangian $\boldsymbol{\mathcal{L}}_{ck}^{X,X}$, and the ESP Lagrangian $\boldsymbol{\mathcal{L}}_{ck}^{X,K}$ are defined as

$$\widetilde{\boldsymbol{\mathcal{L}}}_{ck}^{X,X}=4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}P_{Ii}^{CA}A_{Ii,ck}^{X,X}-2\sum_{i'i\in X}^{actX}P_{i'i}^{AA}A_{i'i,ck}^{X,X}+2\sum_{a'a\in X}^{virtX}P_{a'a}^{VV}A_{a'a,ck}^{X,X}+4\delta_{kAct}L_{ck}-4\widetilde{L}_{kc} \tag{62}$$

$$\overline{\boldsymbol{\mathcal{L}}}_{ck}^{X,K}=4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}P_{Ii}^{CA}A_{Ii,ck}^{X,K}-2\sum_{i'i\in X}^{actX}P_{i'i}^{AA}A_{i'i,ck}^{X,K}+2\sum_{a'a\in X}^{virtX}P_{a'a}^{VV}A_{a'a,ck}^{X,K} \tag{63}$$

If fragment X is a dimer, the *dimer* response term $\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X,X}$ needs to be transformed into a *monomer* response term using Eq. (48) as follows

$$\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}U_{ck}^{X,\zeta}\boldsymbol{\mathcal{L}}_{ck}^{X,X}=\sum_{d\in X}^{virtX}\sum_{l\in X}^{occX}B_{dl}^{X,(\zeta)}Z_{dl}^{X,VO}+\sum_{K\neq X}\sum_{c\in K}^{virtK}\sum_{k\in K}^{occK}U_{ck}^{K,\zeta}\sum_{d\in X}^{virtX}\sum_{l\in X}^{occX}Z_{dl}^{X,VO}A_{dl,ck}^{X,K} \tag{64}$$

In Eq. (64), the *dimer* Z-vector is obtained by solving the following Z-vector equation, in which $\tilde{\tilde{A}}^{X,X}$ is defined in Eqs. (41) and (33); and the Lagrangian $\boldsymbol{\mathcal{L}}^{X,X}$ defined in Eq. (62)

$$\overset{\approx}{A}{}^{X,X}Z^{X,VO}=\boldsymbol{\mathcal{L}}^{X,X} \tag{65}$$

Further expanding $B_{pq}^{X,(\zeta)}$ using Eq. (32), and rearranging terms, the fragment correlation energy gradient $\partial E_X^{(2)}/\partial\zeta$ can be split into three distinct parts; they are

i) The internal gradient term $\tilde{E}_X^{(2),\zeta}$, in which $\delta_{XDim} = 1$ when fragment X is a dimer, otherwise it is *zero*.

$$
\begin{aligned}
\tilde{E}_X^{(2),\zeta} &= 2\sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX}\left(ia\,|\,jb\right)^{(\zeta)} T_{ab}^{ij} \\
&+\left(\begin{array}{l} 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}\tilde{F}_{Ii}^{X,(\zeta)} P_{Ii}^{X,CA} - 2\sum_{i'i\in X}^{actX}\tilde{F}_{i'i}^{X,(\zeta)} P_{i'i}^{X,AA} \\ +2\sum_{a'a\in X}^{virtX}\tilde{F}_{a'a}^{X,(\zeta)} P_{a'a}^{X,VV} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}\tilde{F}_{ck}^{X,(\zeta)} Z_{ck}^{X,VO} \end{array}\right) \\
&-\left(2\sum_{i'i\in X}^{actX} S_{i'i}^{X,(\zeta)} L_{i'i}^{X} + 4\sum_{k\in X}^{occX}\sum_{c\in X}^{virtX} S_{kc}^{X,(\zeta)}\tilde{L}_{kc}^{X} + 2\sum_{a'a\in X}^{virtX} S_{a'a}^{X,(\zeta)}\tilde{L}_{a'a}^{X}\right) \\
&-\left(\begin{array}{l} 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} S_{Ii}^{X,(\zeta)}\varepsilon_{Ii}^{X} P_{Ii}^{X,CA} - 2\sum_{i'i\in X}^{actX} S_{i'i}^{X,(\zeta)}\varepsilon_{i'i}^{X} P_{i'i}^{X,AA} \\ +2\sum_{a'a\in X}^{virtX} S_{a'a}^{X,(\zeta)}\varepsilon_{a'a}^{X} P_{a'a}^{X,VV} - \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} S_{ck}^{X,(\zeta)}\varepsilon_{k}^{X} Z_{ck}^{X,VO} \end{array}\right) \\
&-\sum_{kl\in X}^{occX} S_{kl}^{X,(\zeta)}\left[\begin{array}{l} 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} P_{Ii}^{X,CA}\tilde{A}_{Ii,kl}^{X,X} - 2\sum_{i'i\in X}^{actX} P_{i'i}^{X,AA}\tilde{A}_{i'i,kl}^{X,X} \\ +2\sum_{a'a\in X}^{virtX} P_{a'a}^{X,VV}\tilde{A}_{a'a,kl}^{X,X} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k'\in X}^{occX} Z_{ck'}^{X,VO}\tilde{A}_{ck',kl}^{X,X} \end{array}\right]
\end{aligned}
\tag{66}
$$

ii) The ESP gradient term $\bar{E}_X^{(2),\zeta}$

$$
\begin{aligned}
\overline{E}_X^{(2),\zeta} &= 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX}\overline{F}_{Ii}^{X,(\zeta)} P_{Ii}^{X,CA} - 2\sum_{i'i\in X}^{actX}\overline{F}_{i'i}^{X,(\zeta)} P_{i'i}^{X,AA} \\
&+2\sum_{a'a\in X}^{virtX}\overline{F}_{a'a}^{X,(\zeta)} P_{a'a}^{X,VV} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX}\overline{F}_{ck}^{X,(\zeta)} Z_{ck}^{X,VO} \\
&-\frac{1}{2}\sum_{K\neq X}\sum_{kl\in K}^{occK} S_{kl}^{K,(\zeta)}\left[\begin{array}{l} 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} P_{Ii}^{X,CA} A_{Ii,kl}^{X,K} - 2\sum_{i'i\in X}^{actX} P_{i'i}^{X,AA} A_{i'i,kl}^{X,K} \\ +2\sum_{a'a\in X}^{virtX} P_{a'a}^{X,VV} A_{a'a,kl}^{X,K} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k'\in X}^{occX} Z_{ck'}^{X,VO} A_{ck',kl}^{X,K} \end{array}\right]
\end{aligned}
\tag{67}
$$

iii) and the collective monomer response terms $U_X^{(2),\zeta}$, in which $\delta_{XMon} = 1$ if fragment X is a monomer.

$$U_X^{(2),\zeta} = \sum_K \sum_{c\in K}^{virtK} \sum_{k\in K}^{occK} U_{ck}^{K,\zeta} \left[\delta_{XMon}\delta_{XK} \widetilde{\boldsymbol{\mathcal{L}}}_{ck}^{X,K} + \left(1-\delta_{XK}\right) \overline{\overline{\boldsymbol{\mathcal{L}}}}_{ck}^{X,K} \right] \tag{68}$$

$$\overline{\overline{\boldsymbol{\mathcal{L}}}}_{ck}^{X,K} = 4\sum_{I\in X}^{corX} \sum_{i\in X}^{actX} P_{Ii}^{CA} A_{Ii,ck}^{X,K} - 2\sum_{i'i\in X}^{actX} P_{i'i}^{AA} A_{i'i,ck}^{X,K} + 2\sum_{a'a\in X}^{virtX} P_{a'a}^{VV} A_{a'a,ck}^{X,K} + \delta_{XDim} \sum_{d\in X}^{virtX} \sum_{l\in X}^{occX} Z_{dl}^{X,VO} A_{dl,ck}^{X,K} \tag{69}$$

In GAMESS, the internal fragment MP2 gradient term $\tilde{E}_X^{(2),\zeta}$ is evaluated in the MP2 driver with a little modification compared with the regular MP2 driver. The major difference is that after the internal fragment Lagrangian $\boldsymbol{\mathcal{L}}^{X,X}$ is calculated (Eq. (62)), if fragment X is a *monomer* ($\delta_{XMon} = 1$), the internal fragment Lagrangian $\boldsymbol{\mathcal{L}}^{X,X}$ is passed to the SCZV solver. If the fragment X is a *dimer* ($\delta_{XDi} = 1$), the Z-vector equation is solved for the *dimer* Z-vector (Eq. (65)), followed by the transformation of the dimer response term into the *monomer* response term (Eq. (64)). Note that the *dimer* Z-vector is not directly passed to the SCZV solver as shown in Eq. (64) but through the MP2 correlation density discussed in the next section.

4.3.2 The internal fragment gradient

The MP2 internal fragment gradient can be simplified by i) expanding the internal Fock matrix element $\tilde{F}_{pq}^X$ defined in Eq. (52); ii) transforming the MP2 density $P^{X,CA}$, $P^{X,AA}$, $P^{X,VV}$ and $Z^{X,VO}$ [if fragment X is a *dimer*] from the MO to the AO basis; iii) transforming the energy-weighted density from the MO to the AO basis. The MP2 density is given as follows

$$
\begin{aligned}
D_{\mu\nu}^{X,(2)} = & 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} C_{\mu I}^{X} C_{\nu i}^{X} P_{Ii}^{X,CA} - 2\sum_{i'i\in X}^{actX} C_{\mu i'}^{X} C_{\nu i}^{X} P_{i'i}^{X,AA} \\
& +2\sum_{a'a\in X}^{virtX} C_{\mu a'}^{X} C_{\nu a}^{X} P_{a'a}^{X,VV} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} C_{\mu c}^{X} C_{\nu k}^{X} Z_{ck}^{X,VO}
\end{aligned}
\tag{70}
$$

The MP2 energy-weighted density $\widetilde{W}_{\mu\nu}^{X,(2)}$ consists of three components. In Eq. (73), the average pair-wise MO energy ε_{pq}^{X} is defined as $\left(\varepsilon_{p}^{X} + \varepsilon_{q}^{X}\right)/2$.

$$
\widetilde{W}_{\mu\nu}^{X,(2)} = \widetilde{W}[I]_{\mu\nu}^{X,(2)} + \widetilde{W}[II]_{\mu\nu}^{X,(2)} + \widetilde{W}[III]_{\mu\nu}^{X,(2)} \tag{71}
$$

$$
\widetilde{W}[I]_{\mu\nu}^{X,(2)} = -2\sum_{i'i\in X}^{actX} C_{\mu i'}^{X} C_{\nu i}^{X} L_{i'i} - 4\sum_{k\in X}^{occX}\sum_{c\in X}^{virtX} C_{\mu k}^{X} C_{\nu c}^{X} \widetilde{L}_{kc} - 2\sum_{a'a\in X}^{virtX} C_{\mu a'}^{X} C_{\nu a}^{X} \widetilde{L}_{a'a} \tag{72}
$$

$$
\begin{aligned}
\widetilde{W}[II]_{\mu\nu}^{X,(2)} = & -4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} C_{\mu I}^{X} C_{\nu i}^{X} \varepsilon_{Ii}^{X} P_{Ii}^{CA} + 2\sum_{i'i\in X}^{actX} C_{\mu i'}^{X} C_{\nu i}^{X} \varepsilon_{i'i}^{X} P_{i'i}^{AA} \\
& -2\sum_{a'a\in X}^{virtX} C_{\mu a'}^{X} C_{\nu a}^{X} \varepsilon_{a'a}^{X} P_{a'a}^{VV} + \delta_{XDim}\sum_{c\in X}^{virtX}\sum_{k\in X}^{occX} C_{\mu c}^{X} C_{\nu k}^{X} \varepsilon_{k}^{X} Z_{ck}^{VO}
\end{aligned}
\tag{73}
$$

$$
\widetilde{W}[III]_{\mu\nu}^{X,(2)} = -\sum_{kl\in X}^{occX} C_{\mu k}^{X} C_{\nu l}^{X} \widetilde{A}_{\lambda\sigma,kl}^{X,X} D_{\lambda\sigma}^{X,(2)} \tag{74}
$$

The MP2 internal fragment gradient term $\tilde{E}_{X}^{(2),\zeta}$ can be rewritten in a neat form using the MP2 correlation density $D_{\mu\nu}^{X,(2)}$, the SCF density $D_{\mu\nu}^{X}$ and the internal correlation energy-weighted density $\widetilde{W}_{\mu\nu}^{X,(2)}$ as follows

$$
\begin{aligned}
\widetilde{E}_{X}^{(2),\zeta} = & 2\sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX} \left(ia\,|\,jb\right)^{(\zeta)} T_{ab}^{ij} \\
& + \sum_{\mu\nu\lambda\sigma\in X}^{aoX} \left(\mu\nu\,|\,\lambda\sigma\right)^{\zeta}\left[D_{\mu\nu}^{X,(2)} D_{\lambda\sigma}^{X} - \frac{1}{2} D_{\mu\lambda}^{X,(2)} D_{\nu\sigma}^{X} \right] \\
& + \sum_{\mu\nu\in X}^{aoX} S_{\mu\nu}^{X,\zeta} \widetilde{W}_{\mu\nu}^{X,(2)} + \sum_{\mu\nu\in X}^{aoX} H_{\mu\nu}^{X,\zeta} D_{\mu\nu}^{X,(2)}
\end{aligned}
\tag{75}
$$

4.3.3 The ESP fragment gradient term

In a similar manner to the internal fragment gradient, the MP2 ESP fragment gradient term $\bar{E}_X^{(2),\zeta}$ can be manipulated by i) expanding the ESP Fock matrix element $\bar{F}_{pq}^X$ (Eq. (57)); and ii) defining the ESP energy-weighted density $\bar{W}_{\mu\nu}^K$ using the MP2 density as follows

$$\bar{W}_{\mu\nu}^K = -\frac{1}{2}\sum_{kl\in K}^{occK}\sum_{\lambda\sigma\in X}^{aoX} C_{\mu k}^K C_{\nu l}^K A_{\lambda\sigma,\mu\nu}^{X,K} D_{\lambda\sigma}^{X,(2)} \tag{76}$$

The ESP fragment gradient term becomes, in which $\bar{u}_{\mu\nu}^{X,\zeta}$ is the electrostatic interaction between electrons in fragment X and nuclei in the other fragments, which is defined in Eq. (58)

$$\bar{E}_X^{(2),\zeta} = \sum_{K\neq X}\sum_{\mu\nu\in K}^{aoK} S_{\mu\nu}^{X,\zeta}\bar{W}_{\mu\nu}^K + \sum_{\mu\nu\in X}^{aoX}\bar{u}_{\mu\nu}^{X,\zeta} D_{\mu\nu}^{X,(2)} + \frac{1}{2}\sum_{K\neq X}\sum_{\lambda\sigma\in K}^{aoK}\sum_{\mu\nu\in X}^{occX}(\mu\nu\,|\,\lambda\sigma)^{\zeta} D_{\mu\nu}^{X,(2)} D_{\lambda\sigma}^{K} \tag{77}$$

4.3.4 Collective monomer response terms

Finally, in terms of the MP2 density, the ESP Lagrangian $\bar{\bar{\mathcal{L}}}_{ck}^{X,K}$ in Eq. (69) can be simplified as

$$\bar{\bar{\mathcal{L}}}_{ck}^{X,K} = \sum_{\lambda\sigma\in X}^{aoX} A_{\lambda\sigma,ck}^{X,K} D_{\lambda\sigma}^{X,(2)} \tag{78}$$

4.4 FMO2/RI-MP2 analytic gradient

To reduce the computational cost, the RI approximation is applied to the internal fragment gradient term $\tilde{E}_X^{(2),\zeta}$ [Eq. (75)]. This includes the evaluation of the density matrices, the internal fragment Lagrangian, and the 4-2ERI derivative in terms of the RI approximation. The ESP gradient term $\bar{E}_X^{(2),\zeta}$ [Eq. (77)] and the response terms $U^{(2),\zeta}$ [Eq. (68)] also inherit benefits from the RI approximation since their main inputs are density matrices form the internal fragment gradient terms.

4.4.1 RI approximation

In the RI approximation, a 4-2ERI of a fragment *X* is approximated by the product of 3- and 2-2ERIs as follows

$$(\mu\nu\,|\,\lambda\sigma) \simeq \sum_{PQ\in X}^{auxX} (\mu\nu\,|\,P) V_{PQ}^{X,-1} (Q\,|\,\lambda\sigma) \tag{79}$$

$$(\mu\nu\,|\,P) = \iint dr_1 dr_2 \phi_\mu^*(r_1)\phi_\nu(r_1) r_{12}^{-1} \alpha_P(r_2) \tag{80}$$

$$V_{PQ}^X = \iint dr_1 dr_2 \alpha_P(r_1) r_{12}^{-1} \alpha_Q(r_2) \tag{81}$$

$auxX$ is the auxiliary basis of the fragment X. The inverse of the matrix V^X can be decomposed and combined with the 3-2ERIs to form the 3-index matrix $\tilde{B}[X]$ that can be used to form 4-2ERIs on-the-fly

$$V_{PQ}^{X,-1} = \sum_{R\in X}^{auxX} \Omega_{PR}^X \Omega_{RQ}^{X,\dagger} \tag{82}$$

$$\widetilde{B}[X]_{\mu\nu}^P = \sum_{R\in X}^{auxX} (\mu\nu\,|\,R) \Omega_{RP}^X \tag{83}$$

$$(\mu\nu\,|\,\lambda\sigma) = \sum_{P\in X}^{auxX} \widetilde{B}[X]_{\mu\nu}^P \widetilde{B}[X]_{\lambda\sigma}^{P,\dagger} \tag{84}$$

The matrix $\tilde{B}[X]$ can also be transformed into the matrix $B[X]$ in the MO basis (Eq. (73)), which can be used to form 4-2ERIs in the MO basis (Eq. (74)). As only two AO indices need to be transformed, this introduces the main computational savings of the RI approximation (e.g., in correlation methods that need 4-2ERIs in the MO basis)

$$B[X]_{pq}^P = \sum_{\mu\nu\in X}^{aoX} C_{\mu p}^X C_{\nu q}^X \widetilde{B}[X]_{\mu\nu}^P \tag{85}$$

$$(pq|rs) = \sum_{P\in X}^{auxX} B[X]_{pq}^{P} B[X]_{rs}^{P,\dagger} \tag{86}$$

The derivatives of the RI 4-2ERIs can be expanded in terms of the derivatives of 3- and 2-ERIs, whose analytic and closed forms are:

$$\begin{aligned}(\mu\nu|\lambda\sigma)^{\zeta} \simeq & \sum_{PQ\in X}^{auxX}\left[(\mu\nu|P)^{\zeta}\,\Omega_{PQ}^{X} B[X]_{\lambda\sigma}^{Q,\dagger} + B[X]_{\mu\nu}^{P}\,\Omega_{PQ}^{X,\dagger}(Q|\lambda\sigma)^{\zeta}\right] \\ & - \sum_{PQRS\in X}^{auxX} B[X]_{\mu\nu}^{R}\,\Omega_{RP}^{X,\dagger} V_{PQ}^{X,\zeta}\Omega_{QS}^{X} B[X]_{\lambda\sigma}^{S,\dagger}\end{aligned} \tag{87}$$

4.4.2 FMO/RI-MP2 internal fragment gradient term

For the internal fragment gradient term $\tilde{E}_X^{(2),\zeta}$ [Eq. (75)], the RI approximation [Eqs. (79)-(87)] can be used to evaluate the MP2 amplitude T_{ab}^{ij} [Eq. (52)], 2-particle density matrices (2PDM) $P_{Ii}^{X,CA}, P_{i'i}^{X,AA}, P_{a'a}^{X,VV}$ [Eqs. (58)-(60)], $L_{ci}^{X}, \tilde{L}_{kc}^{X}$ [Eqs. (56), (57)] the internal Lagrangian $\boldsymbol{\mathcal{L}}_{ck}^{X,X}$ [Eq. (62)], and the first 4-2ERI derivative term $2\sum_{ij\in X}^{actX}\sum_{ab\in X}^{virtX}(ia|jb)^{(\zeta)}T_{ab}^{ij}$ [the first term in Eq. (75)]. For clarity, the subscript or superscript X that stands for the general fragment X is dropped in this section. First, applying the RI approximation, the internal fragment gradient becomes

$$\begin{aligned}\frac{\partial}{\partial\zeta}\widetilde{E}_X^{(2)} \simeq & \sum_{\mu\nu\in X}^{aoX}\sum_{P\in X}^{auxX}(\mu\nu|P)^{\zeta}\,\widetilde{\Gamma}_{\mu\nu}^{P} + \sum_{PQ\in X}^{auxX} V_{PQ}^{\zeta}\gamma_{PQ} \\ & + \sum_{\mu\nu\lambda\sigma\in X}^{aoX}(\mu\nu|\lambda\sigma)^{\zeta}\left[D_{\mu\nu}^{X,(2)}D_{\lambda\sigma}^{X} - \frac{1}{2}D_{\mu\lambda}^{X,(2)}D_{\nu\sigma}^{X}\right] \\ & + \sum_{\mu\nu\in X}^{aoX} S_{\mu\nu}^{X,\zeta}\widetilde{W}_{\mu\nu}^{X} + \sum_{\mu\nu\in X}^{aoX} H_{\mu\nu}^{X,\zeta}D_{\mu\nu}^{X,(2)}\end{aligned} \tag{88}$$

The first two terms in Eq. (88) arise from the application of the RI approximation to the 4-2ERI derivative. The 3-index 2-particle density matrix (3-2PDM) in the AO basis $\tilde{\Gamma}_{\mu\nu}^{P}$ is defined in Eqs. (89)-(93). The 2-2PDM γ_{PQ} is subsequently obtained from the 3-2PDM in the MO basis Γ_{ia}^{P}

$$\widetilde{\Gamma}^{P}_{\mu\nu} = 4\sum_{i\in X}^{actX}\sum_{a\in X}^{virtX} C^{X}_{\mu i}\Gamma^{P}_{ia}C^{X,\dagger}_{a\nu} \tag{89}$$

$$\Gamma^{P}_{ia} = \sum_{Q\in X}^{auxX} \Omega_{PQ} Y^{Q}_{ia} \tag{90}$$

$$Y^{Q}_{ia} = \sum_{j\in X}^{actX}\sum_{b\in X}^{virtX} B^{Q}_{jb} T^{ij}_{ab} \tag{91}$$

$$T^{ij}_{ab} = 2t^{ij}_{ab} - t^{ij,\dagger}_{ba} \tag{92}$$

$$t^{ij}_{ab} = \sum_{P\in X}^{auxX} B^{P}_{ia} B^{P,\dagger}_{jb} \Big/ D^{ij}_{ab} \tag{93}$$

$$\gamma_{PQ} = \sum_{R\in X}^{auxX}\left(\sum_{i\in X}^{actX}\sum_{a\in X}^{virtX}\Gamma^{Q}_{ia}B^{R,\dagger}_{ia}\right)\Omega_{RP} \tag{94}$$

The MP2 correlation density $D^{X,(2)}_{\mu\nu}$ [Eq. (70)] is formed from the density matrices $P^{CA}_{Ii}, P^{AA}_{i'i}, P^{VV}_{a'a}$ [Eqs. (58)-(60)] and the Z-vector Z^{VO}_{ck} [if the fragment X is a dimer]. The first three matrices along with $L^{X}_{ci}, \tilde{L}^{X}_{kc}$ [Eqs. (56), (57)] are used to build the internal Lagrangian $\mathcal{L}^{X,X}_{ck}$ [Eq. (62)]. For monomers, the Lagrangian is copied to the SCZV solver. For *dimers*, the Z-vector equation is solved for the *dimer* Z-vector Z^{VO}_{ck} that indirectly enters the SCZV solver in the form of the MP2 correlation density matrix $D^{X,(2)}_{\mu\nu}$

$$P^{VV}_{a'a} = \sum_{ij\in X}^{actX}\sum_{b\in X}^{virtX}\frac{\sum_{P\in X}^{auxX} B^{P}_{ia'}B^{P,\dagger}_{jb}}{D^{ij}_{a'b}}T^{ij}_{ab} \tag{95}$$

$$P^{AA}_{i'i} = \sum_{j\in X}^{actX}\sum_{ab\in X}^{virtX}\frac{\sum_{P\in X}^{auxX} B^{P}_{i'a}B^{P,\dagger}_{jb}}{D^{i'j}_{ab}}T^{ij}_{ab} \tag{96}$$

$$L_{mi} \simeq \sum_{a\in X}^{virtX}\sum_{P\in X}^{auxX} B_{ma}^{P,\dagger}\Gamma_{ia}^{P} \tag{97}$$

$$P_{Ii}^{CA} = \frac{L_{Ii}}{e_i^X - e_I^X} \tag{98}$$

$$\widetilde{L}_{ma} \simeq \sum_{i\in X}^{actX}\sum_{P\in X}^{auxX} B_{im}^{P,\dagger}\Gamma_{ia}^{P} \tag{99}$$

$$\mathcal{L}_{ck}^{X,X} = 4\sum_{I\in X}^{corX}\sum_{i\in X}^{actX} A_{Ii,ck}^{X,X}P_{Ii}^{CA} - 2\sum_{i'i\in X}^{actX} A_{i'i,ck}^{X,X}P_{i'i}^{AA} + 2\sum_{a'a\in X}^{virtX} A_{a'a,ck}^{X,X}P_{a'a}^{VV} + 4\delta_{kAct}L_{ck} - 4\widetilde{L}_{kc} \tag{100}$$

In Eq. (100), the $\delta_{kAct} = 1$ only when k is the index of an active occupied MO, otherwise it is zero.

4.4.3 FMO/RI-MP2 gradient: brief summary

The FMO/RI-MP2 gradient is briefly illustrated in FIGURE 4.1. After the SCC procedure and/or dimer SCF calculations, the fragment densities are available and are used as the input for the MP2 fragment gradient evaluation. The internal fragment MP2 gradient is evaluated using the RI approximation. The specific algorithm that applies the RI approximation to evaluate the density matrices and other terms will be discussed in the next sections. Besides the gradient contribution, the output from the internal fragment gradient including the MP2 density $D_{\mu\nu}^{X,(2)}$ and the *monomer* internal Lagrangian $\mathcal{L}_{kc}^{X,X}$ are passed to the ESP fragment gradient and the SCZV drivers to evaluate gradient contributions of the ESP and response terms. For dimers, only the MP2 density is passed to the SCZV driver.

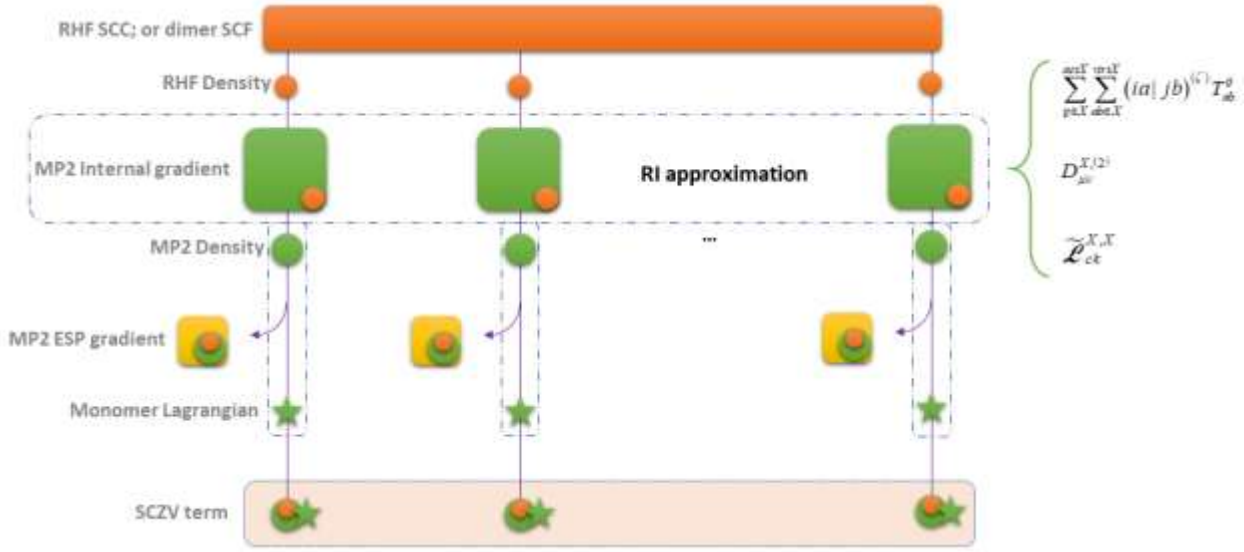


FIGURE 4.1. Workflow of FMO/RI-MP2 gradient.

4.4.4 FMO/RI-MP2 internal fragment gradient term: serial implementation.

A serial implementation for the internal gradient term $\tilde{E}_X^{(2),\zeta}$ using the RI approximation is briefly presented in SCHEME 4.1. The RI approximation is used to evaluate the MP2 amplitude, the density matrices, and the first 4-2ERI derivative term. In SCHEME 4.1, the integrals and/or the matrix elements are usually evaluated and stored in arrays with the dimensions being the number of core (C), active occupied (A), and/or virtual (V) MOs. The sequence of processes in SCHEME 4.1 is:

i) The MP2 amplitude is evaluated for pair-by-pair of active occupied MOs and stored in T_{VV} [lines 3-6]. The fundamental component to form the MP2 amplitude is t_{VV} , which is the ratio of the 4-2ERIs and the active occupied-virtual pairwise MO energy difference D_{VV}^{ij}.

ii) After the MP2 amplitude T_{VV} is formed, the density matrix $P^{VV}_{a'a}$ can readily be formed [lines 7-8] using the same t_{VV}. Similarly, the $P^{AA}_{i'i}$ can also be calculated; nevertheless, this needs the 4-2ERIs to be re-calculated [lines 9-14].

iii) The intermediate Y^{Q}_{ia} (Eq. (91)) is evaluated and stored in Y^{i}_{XV} [line 15-16], which is then used to form the 3-2PDM Γ^{P}_{ia} (Eq. (90)) and stored in Γ^{i}_{XV} [line 18-19]. The 3-2PDM Γ^{i}_{XV} is then used to as the intermediate to form P^{CA}_{Ii}, L_{ci}, $\tilde{L}_{kc}$ and $\tilde{L}_{a'a}$ defined in Eqs. (97)-(99). These matrix elements are calculated and stored in the arrays P^{CA}_{CA} [lines 20-24], L_{VA} [lines 25-28], $\tilde{L}_{OV}$ [lines 29-30] and $\tilde{L}_{VV}$ [lines 31-32], respectively. The 3-2PDM Γ^{i}_{XV} is also used to form $\tilde{\gamma}_{XX}$ [lines 33-34], which is an intermediate to evaluate the 2-2PDM.

iv) When the active-active loop [lines 1-35] finishes, the 3-2PDM $\Gamma^{P}_{\mu\nu}$ and 2-2PDMs γ_{PQ} [lines 36-44] can be formed and combined with the derivatives of the 3-2ERIs $(\mu\nu|P)^{\zeta}$ and 2-2ERIs $(P|Q)^{\zeta}$ for contributions to the MP2 internal fragment gradient $\tilde{E}^{(2),\zeta}_{X}$.

v) The density matrices $P^{CA}_{Ii}, P^{AA}_{i'i}, P^{VV}_{a'a}$ and $L_{ci}, \tilde{L}_{kc}$ are used to build the internal fragment Lagrangian $\boldsymbol{\mathcal{L}}^{X,X}_{ck}$ (Eq. (100)). $\boldsymbol{\mathcal{L}}^{X,X}_{ck}$ is saved as input for the SCZV solver if the fragment X is a monomer [lines 48-50]; otherwise, it is used to solve for the dimer Z-vector Z^{VO}_{ck} [lines 51-53].

vi) The complete set of $P^{CA}_{Ii}, P^{AA}_{i'i}, P^{VV}_{a'a}$ and Z^{VO}_{ck} [if the fragment is a dimer] is also used to build the energy-weighted density $\widetilde{W}^{X,(2)}_{\mu\nu}$, which is combined with the derivative of the overlap integrals $S^{X,\zeta}_{\mu\nu}$ for another contribution to the MP2 internal fragment gradient [lines 55-57].

vii) The MP2 correlation density in the AO basis $D^{X,(2)}_{\mu\nu}$ is formed from $P^{CA}_{Ii}, P^{AA}_{i'i}, P^{VV}_{a'a}$ and Z^{VO}_{ck} [if the fragment is a dimer] in lines 59-60. The density is saved for the ESP fragment

gradient term and the SCZV solver [lines 62-63]. Finally, the MP2 correlation density $D_{\mu\nu}^{X,(2)}$ is combined with the SCF density $D_{\mu\nu}^{X}$ and the derivative of the 4-2ERI $(\mu\nu|\lambda\sigma)^{\zeta}$ for the final contribution to the MP2 internal fragment gradient [lines 66-68].

SCHEME 4.1. MP2 internal gradient term evaluation using the RI approximation

1 $do\ i\ = 1, nact$
2 $do\ j = 1, nact$
3 //form: T_{ab}^{ij}
4 $t_{VV} = B_{VX}^{i,\dagger} \times B_{XV}^{j}$
5 $t_{VV} = t_{VV}/D_{VV}^{ij}$
6 $T_{VV} = t_{VV} + t_{VV} - t_{VV}^{\dagger}$
7 //form: $P_{a'a}^{VV}$
8 $\tilde{P}_{VV}^{VV} \leftarrow t_{VV} \times T_{VV}$
9 //form: $P_{i'i}^{AA}$
10 $do\ i' = 1, nact$
11 $t_{VV} = B_{VX}^{i',\dagger} \times B_{XV}^{j}$
12 $t_{VV} = t_{VV}/D_{VV}^{i'j}$
13 $P_{i'i}^{AA} \leftarrow t_{VV} \times T_{VV}$
14 $enddo\ i'$
15 //form: Y_{Pa}^{i}
16 $Y_{XV} \leftarrow B_{XV}^{j} \times T_{VV}$
17 $enddo\ j$
18 //form: $\mathbf{\Gamma}_{Pa}^{i}$
19 $\mathbf{\Gamma}_{XV}^{i} = \mathbf{\Omega}_{XX} \times Y_{XV}$
20 //form: P_{Ii}^{CA} from L_{Ii}
21 $do\ I\ \in coreX$
22 $P_{Ii}^{CA} = B_{VX}^{I,\dagger} \cdot \mathbf{\Gamma}_{XV}^{i}$
23 $P_{Ii}^{CA} = P_{Ii}^{CA}/(\boldsymbol{\epsilon}_{i}^{X} - \boldsymbol{\epsilon}_{I}^{X})$
24 $enddo\ I$
25 //form: L_{ci}
26 $do\ c\ \in virtX$
27 $L_{ci} = B_{VX}^{c,\dagger} \cdot \mathbf{\Gamma}_{XV}^{i}$
28 $enddo\ c$
29 //form:$\tilde{L}_{kc}$
30 $\tilde{L}_{OV} \leftarrow B_{OX}^{i,\dagger} \times \mathbf{\Gamma}_{XV}^{i}$
31 //form: $\tilde{L}_{a'a}$
32 $\tilde{L}_{VV} \leftarrow B_{VX}^{i,\dagger} \times \mathbf{\Gamma}_{XV}^{i}$
33 //form: $\tilde{\mathbf{\gamma}}_{XX}$
34 $\tilde{\mathbf{\gamma}}_{XX} \leftarrow \mathbf{\Gamma}_{XV}^{i} \times B_{VX}^{i,\dagger}$
35 $enddo\ i$
36 //2-2ERI grd contribution
37 $\mathbf{\gamma}_{XX} = -2.0 \times \tilde{\mathbf{\gamma}}_{XX} \times \mathbf{\Omega}_{XX}$
38 $\tilde{E}_{X}^{(2),\boldsymbol{\zeta}} \leftarrow (\mathbf{\gamma}_{XX}, V_{XX}^{\boldsymbol{\zeta}})$
39
40 //3-2ERI grd contribution
41 $do\ P \in auxX$
42 $\tilde{\mathbf{\Gamma}}_{NN}^{P} = 4 \times C_{NA}^{X} \mathbf{\Gamma}_{PV}^{A} C_{VN}^{X,\dagger}$
43 $\tilde{E}_{X}^{(2),\boldsymbol{\zeta}} \leftarrow ((\boldsymbol{\mu\nu}|P)^{\boldsymbol{\zeta}}, \tilde{\mathbf{\Gamma}}_{NN}^{P})$
44 $enddo\ P$
45
46 //Lagrangian and Z-vector
47 $\mathcal{L}_{ck}^{X,X} \leftarrow (P_{Ii}^{CA}, P_{i'i}^{AA}, P_{a'a}^{VV}, L_{ck}, \tilde{L}_{kc})$
48 $if(X = MONOMER) then$
49 //Save $\mathcal{L}_{ck}^{X,X}$ for SCZV solver
50 $Save$: $\mathcal{L}_{ck}^{X,X}$
51 $elif(X = DIMER)\ then$
52 $Z_{ck}^{VO} \leftarrow (\mathcal{L}_{ck}^{X,X}, \tilde{A}_{dl,ck}^{X,X})$
53 $endif$
54
55 //Overlap grd contribution
56 $\widetilde{W}_{\boldsymbol{\mu\nu}}^{X,(2)} \leftarrow P_{Ii}^{CA}, P_{i'i}^{AA}, P_{a'a}^{VV}, Z_{ck}^{VO}[DIMER]$
57 $\tilde{E}_{X}^{(2),\boldsymbol{\zeta}} \leftarrow \left(\widetilde{W}_{\boldsymbol{\mu\nu}}^{X,(2)}, S_{\boldsymbol{\mu\nu}}^{X,\boldsymbol{\zeta}}\right)$
58
59 //Form MP2 correlation density
60 $D_{\boldsymbol{\mu\nu}}^{X,(2)} \leftarrow P_{Ii}^{CA}, P_{i'i}^{AA}, P_{a'a}^{VV}, Z_{ck}^{VO}[DIMER]$
61
62 //Save $D_{\boldsymbol{\mu\nu}}^{X,(2)}$ for ESP term and SCZV solver
63 $Save$: $D_{\boldsymbol{\mu\nu}}^{X,(2)}$
64
65 //4-2ERI grd contribution
66 $D_{\boldsymbol{\mu\nu}}^{X,(2)} \leftarrow P_{Ii}^{CA}, P_{i'i}^{AA}, P_{a'a}^{VV}, Z_{ck}^{VO}[DIMER]$
67 $\tilde{E}_{X}^{(2),\boldsymbol{\zeta}} \leftarrow \left((\boldsymbol{\mu\nu}|\boldsymbol{\lambda\sigma})^{\boldsymbol{\zeta}}, D_{\boldsymbol{\mu\nu}}^{X,(2)}, D_{\boldsymbol{\mu\nu}}^{SCF}\right)$

4.4.5 FMO/RI-MP2 internal fragment gradient term with GDDI.

In GAMESS, the FMO methods are supported by the GDDI[7] approach to parallelism that can split compute processes [ranks] into groups. Within the scope of each group, ranks can send and receive messages to each other. A global sum of a floating point or integer variable or array can also be done within a group scope. The GDDI, therefore, facilitates each group of ranks to work on a chunk of work [e.g., a fragment] relatively independently. In the FMO framework, this avoids the need for all ranks to work on one small fragment at a time. An even distribution among fragments accomplishes the node linear scaling of FMO.

Nevertheless, since it is based on a distributed memory model, the GDDI inherits expensive communication overhead and memory footprint that restricts computational performance. For the new hardware generations of multicore processors, a recently introduced hybrid distributed/shared memory GDDI/OpenMP model[22,23] can both preserve the FMO linear scaling and significantly enhance the efficiency of the calculations by reducing the communication overhead and memory footprint.

Since the GDDI[7] has been well documented,[6] this section focuses on the OpenMP implementation for the MP2 internal fragment gradient term with the RI approximation, particularly for the density matrix formation. In the hybrid GDDI/OpenMP model,[22,23] only a small number of ranks are kicked off from each [logical] compute node. These ranks are then split into groups using the GDDI. Each group of $nprocs$ ranks will be assigned a chunk of work; e.g., evaluating density matrices in the MP2 internal fragment gradient term.

In SCHEME 4.2, the active-active loop is distributed among $nprocs$ ranks of a GDDI group by assigning each rank a chunk of loop $(Lstart, Lend)$ as shown in line 6. Subsequently, each rank spawns a team of threads that do the actual computation for the inner loop [lines 7-42]. Within the FMO framework supported by the OpenMP shared memory model, the entire 3-index matrix B (Eq. (73)) can be loaded into the process memory (line 10). The decomposed matrix Ω (Eq. (70)) can also be used as a shared array. Some small intermediate matrices (e.g., t_{VV}, T_{VV}, Y_{XV}) are made private to threads (line 11).

Data accumulation for $P_{AA}^{AA}, P_{VV}^{VV}, Y_{XV}$ are likely to encounter race conditions. This potential problem is addressed by using $threadprivate$ buffers for $\tilde{P}_{AA}^{AA}, \tilde{P}_{VV}^{VV}, \tilde{Y}_{XV}$ in the places that allow data accumulation efficiently during the calculations in the threading region.[22] When the entire loop finishes, the $threadprivate$ arrays are reduced to shared arrays Y_{XV} [lines 27-28] and P_{AA}^{AA}, P_{VV}^{VV} [lines 49-58] in a separate threading OpenMP region only once.

Note that by placing the OpenMP region [lines 7-42] inside a loop, teams of threads are repeatedly created and destroyed. This might introduce thread creation-destroy overhead. This overhead can be removed by setting the wait policy of threads to be active; e.g., *omp_wait_policy*(*active*). Further, in order to keep $threadprivate$ data consistent due to the discontinuity of threading regions, the dynamic thread property needs be turned off; e.g., by calling $omp\_set\_dynamic(false)$.

Finally, the data in $nprocs$ ranks of density matrices $P_{CA}^{CA}, P_{AA}^{AA}, P_{VV}^{VV}$ need to be reduced [summed up, lines 60-63]. Note that while this is done in a manner that is similar to that of regular

DDI calls,[32] without the GDDI support such reduction calls will reduce data of all ranks, not just ranks within the scope of a group.

SCHEME 4.2. OpenMP implementation of density matrices in internal gradient term

```
1  //create threadprivate arrays
2  !$omp threadprivate(P̃^AA_i'i, P̃^VV_VV)
3  //turn off dynamic threads
4  CALL omp_set_dynamic(.F.)
5  //GDDI work distribution
6  do i = Lstart, Lend
7  //start omp region
8  !$OMP PARALLEL
9  //shared and private arrays
10 !$omp shared(B(X,V,A), Ω_XX)
11 !$omp private(t_VV, T_VV, Y_XV)
12 !$omp do
13       do j = 1, nact
14            //form: T^ij_ab
15            t_VV = B^i,†_VX × B^j_XV
16            t_VV = t_VV / D^ij_VV
17            T_VV = t_VV + t_VV − t^†_VV
18            //form: P^VV_a'a
19            P̃^VV_VV ← t_VV × T_VV
20            //form: P^AA_i'i
21            do i' = 1, nact
22                 t_VV = B^i',†_VX × B^j_XV
23                 t_VV = t_VV / D^i'j_VV
24                 P^AA_i'i ← t_VV × T_VV
25            enddo i'
26            //form: Y^Q_ia
27            Ỹ_XV ← B^j_XV × T^ij_VV
28       enddo j
29 !$omp end do
30 //reduce: Y_XV
31 !$omp critical
32       Y_XV ← Ỹ_XV
33 !$omp end critical
34 //form Γ^i_Pa
35 !$omp do
36       do P ∈ auxX
37            Γ^i_PV = Ω_PX × Y_XV
38       enddo P
39 !$omp end do
40 //form: P^CA_Ii from L_Ii
41 !$omp do
42       do I ∈ corX
43            P^CA_Ii = B^I,†_VX · Γ^i_XV
44            P^CA_Ii = P^CA_Ii / (ε^X_i − ε^X_I)
45       enddo I
46 !$omp end do
47 !$OMP END PARALLEL
48 enddo i
49 //OMP reduce density matrices
50 !$OMP PARALLEL
51 !$omp critical
52       P^VV_VV ← P̃^VV_VV
53 !$omp end critical
54
55 !$omp critical
56       P^AA_AA ← P̃^AA_AA
57 !$omp end critical
58 !$OMP END PARALLEL
59
60 //GDDI reduce density matrices
61 GDDI_reduce: P^VV_VV
62 GDDI_reduce: P^AA_AA
63 GDDI_reduce: P^CA_CA
```

4.5 Computational models

In the following sections, test calculations are designed to i) verify that the FMO/RI-MP2 gradient implementation is fully analytic; ii) examine the accuracy and performance of the FMO/RI-MP2 gradient relative to the full FMO/MP2 gradient; and iii) observe the scalability of the hybrid parallel model for the FMO/RI-MP2 gradient across multiple compute nodes.

The testing calculations are carried out on a sequence of water clusters of varying sizes, from 16-2615 molecules. Unless otherwise noted, the number of fragments is set equal to the number of water molecules in the cluster (i.e., one water molecule per fragment). The AO basis set used is 6-31G(d,p). For the RI approximation, the auxiliary basis set can vary from the augmented correlation consistent double to triple zeta basis sets. All calculations are done on the OLCF KNL cluster Theta and the NERSC KNL cluster Cori. Each physical KNL node is split into four logical nodes; for the 64-core KNL nodes on Theta,[33] each logical node has 16 cores; for the 68-core KNL node on Cori,[34] each logical node has 17 cores. For GDDI/OpenMP FMO/RI-MP2 calculations, one rank is created on each logical node, which then spawns a team of threads equal to the number of cores in the logical node (e.g., 16 on Theta, 17 on Cori). By default, all calculations are done with the analytic gradient (e.g., including the SCZV term). The calculated results are compared with the corresponding calculations using the full FMO/MP2 gradient method.

4.6 Results and discussion

4.6.1 Analytic gradient verification

The FMO/RI-MP2 analytic gradient implementation (e.g., with response term included) is verified by initially comparing with the FMO/RI-MP2 numerical gradient. Calculations on water

clusters of 16-64 molecules show that the maximum gradient difference is in the range of $10^{-6} - 10^{-5}$ Hartree/Bohr. For a gradient that is not fully analytic, even a small error can cause serious problems for (for example) molecular dynamics (MD) simulations due to a biased error accumulation.[35] In this section, MD simulations are used to further validate that the FMO/RI-MP2 gradient implementation is fully analytic. Due to the additional response term from the MP2 correlation part, the percent of response term contribution to the total FMO/MP2 gradient is much larger than that of the FMO/RHF gradient.[26] Therefore, introducing approximations might impact the FMO/MP2 analytic gradient accuracy. In the MD simulation, the analytic gradient can be verified in terms of the energy conservation of NVE ensembles. The procedure for checking energy conservation (e.g., for water clusters) using NVE ensembles has been well described in the literature.[10,35] The procedure includes *three* main steps: i) generating the initial geometry of a water cluster; ii) equilibrating the system using an NVT ensemble; and iii) checking the energy conservation in the NVE ensemble. The energy conservation is examined by checking whether the slope of the log-log plot of the root mean square deviation of the energy (RMSD(E)) versus the time step (Δt) is close to 2.0.[10,35] For comparison, the MD simulations are done for both FMO/RI-MP2 and full FMO/MP2 without and with the response terms. The specific steps are:

i) A water cluster of 16 molecules is generated in a box so that the density is ~ the density of water at 300.0 K. This is followed by Monte-Carlo simulations to generate two random initial geometries for the MD simulations.
ii) The water clusters are then equilibrated for 6.0 ps with an NVT MD simulation at 300.0K using the EFP method, using a time step of 1.0 fs. The temperature is regulated by a Nosé-Hoover thermostat, and is rescaled every 1000 fs. This is followed by a 500 fs NVT

MD simulation at 300.0K using the FMO/(RI-)MP2 gradients, with a 1.0 fs time step. The temperature is rescaled every 100.0 fs.

iii) Final geometries and velocities of the NVT equilibrated water clusters are read to the NVE MD simulation using FMO/(RI-)MP2 gradients for 500 steps each. The time step (Δt) is varied from 0.1-1.5 fs. The log-log plots of the RMSD(E) versus time step (Δt) in the NVE ensemble are shown in FIGURE 4.2.

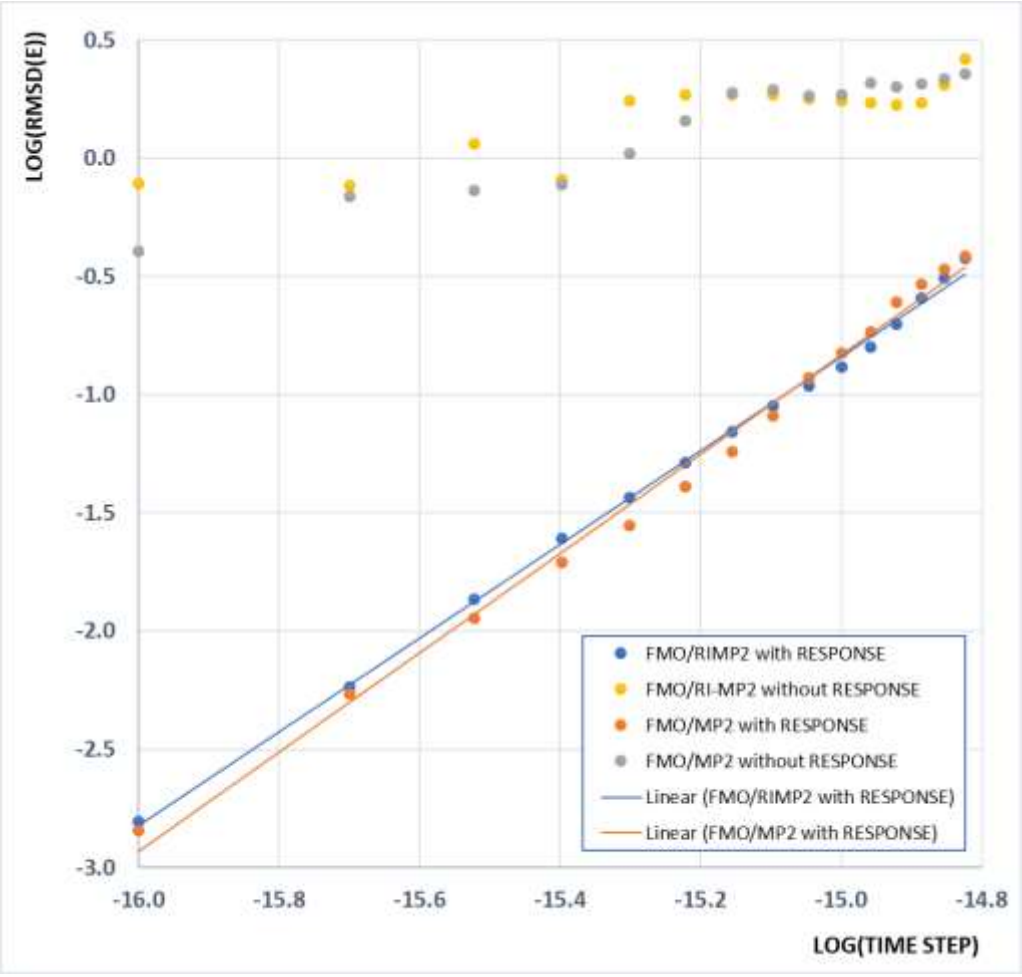


FIGURE 4.2. Log-log plots of the root-mean squared deviation of energy (RMSD(E)) versus time step Δt (0.1-1.5 fs) in NVE ensembles of $(H_2O)_{16}$ clusters using FMO/(RI-)MP2 gradient with and without response terms.

The results show that for randomly different initial geometries, the log-log plots of the NVE MD simulations using both FMO/MP2 and FMO/RI-MP2 analytic gradients, with response terms included, are very close to the ideal slope of 2.0 (e.g., 1.98 and 2.10, respectively), with correlation coefficients R^2 close to 1.0000 (e.g., 0.9979 and 0.9931, respectively). Therefore, the FMO/(RI-)MP2 energy and analytic gradient with response terms included are adequate for the MD simulations. FIGURE 4.2 also shows that NVE ensembles using the FMO/(RI-)MP2 gradient without response terms can cause very large errors in MD simulations as predicted by Nagata et *al*.[26]

4.6.2 Relative accuracy

The accuracy of the FMO/RI-MP2 gradient relative to the full FMO/MP2 gradient is examined using calculations on water clusters of 16-128 molecules. To evaluate the effect of the auxiliary basis in the RI approximation, in all calculations, the AO basis set is fixed to the 6-31G(d,p) basis set, while the auxiliary basis set is varied from augmented correlation consistent double to triple-zeta basis sets. The relative accuracy is assessed using i) the correlation energy error $\Delta E^{(2)}$, which is the absolute correlation energy difference in kcal/mol between the FMO/RI-MP2 and full FMO/MP2 methods; ii) the maximum gradient error $(\Delta g)_{max}$, which is the *maximum* gradient difference in Hartree/Bohr between the two methods; and iii) the root-mean-square gradient deviation $RMSD(g)$ between the two methods, which can be interpreted as the gradient error radius.

TABLE 4.1 shows that the correlation energy error $\Delta E^{(2)}$ is very small, below 1.00 kcal/mol for all water cluster calculations. The energy accuracy is further improved when increasing the

size of the auxiliary basis set. For instance, for $(H_2O)_{64}$ calculations, the energy error decreases from 0.823 to 0.001 kcal/mol when increasing the auxiliary basis set from cc-pVDZ-RI to aug-cc-pVTZ-RI. Similarly, the maximum gradient error $(\Delta g)_{max}$ is also very small, in the range of $10^{-6} - 10^{-5}$ Hartree/Bohr with the error radius $RMSD(g)$ also in the range of $10^{-6} - 10^{-5}$ Hartree/Bohr. The increase in auxiliary basis set also further improves the gradient accuracy. For instance, for the $(H_2O)_{64}$ calculations, the maximum gradient error decreases from 7.73×10^{-5} to 0.6×10^{-5} Hartree/Bohr when increasing the auxiliary basis set from cc-pVDZ-RI to aug-cc-pVTZ-RI.

TABLE 4.1 Accuracy relative to full FMO/MP2 gradient

Water Cluster	Aux. Basis	$\boldsymbol{\Delta E^{(2)}}$ (kcal/mol)	$\boldsymbol{(\Delta g)_{max} \times 10^5}$ (Hartree/Bohr)	$\boldsymbol{RMSD(g) \times 10^5}$ (Hatree/Bohr)
$\boldsymbol{(H_2O)_{16}}$	cc-pVDZ-RI	0.219	3.23	1.55
	aug-cc-pVDZ-RI	0.110	0.94	4.15
	cc-pVTZ-RI	0.007	0.22	0.93
	aug-cc-pVTZ-RI	0.000	0.08	0.38
$\boldsymbol{(H_2O)_{32}}$	cc-pVDZ-RI	0.464	3.85	1.73
	aug-cc-pVDZ-RI	0.239	1.06	0.38
	cc-pVTZ-RI	0.012	0.43	0.16
	aug-cc-pVTZ-RI	0.001	0.27	0.10
$\boldsymbol{(H_2O)_{64}}$	cc-pVDZ-RI	0.823	7.73	2.75
	aug-cc-pVDZ-RI	0.367	1.40	0.58
	cc-pVTZ-RI	0.005	0.89	0.33
	aug-cc-pVTZ-RI	0.001	0.60	0.24
$\boldsymbol{(H_2O)_{128}}$	cc-pVDZ-RI	0.697	7.16	5.87
	aug-cc-pVDZ-RI	0.617	3.06	1.37
	cc-pVTZ-RI	0.617	3.06	1.37
	aug-cc-pVTZ-RI	0.012	1.23	0.47

4.6.3 Relative performance

The performance of the GDDI/OpenMP FMO/RI-MP2 gradient relative to the GDDI FMO/MP2 gradient is defined as the ratio (speedup) of the wall times for the correlation gradient part of the two methods. The wall time and speedup for water clusters of 64 and 128 molecules for

calculations using 2-8 KNL compute nodes on the Cori cluster[34] are presented in TABLE 2. The overall speedup of all calculations is 3.9-8.4x. For a reasonable choice of AO/auxiliary basis set (e.g., 6-31G(d,p)/cc-pVDZ-RI) with an energy accuracy of about 1.0 kcal/mol and gradient difference of about 10^{-5} Hartree/Bohr, the speedup is in the range of 6.2-8.4x. For a larger auxiliary basis set (e.g., 6-31G(d,p)/aug-cc-pVTZ-RI), the speedup is about 4.3-6.7x. Note that an increase in the number of nodes does increase the speedup. For instance, for calculations on the 64-water cluster using the cc-pVDZ-RI auxiliary basis, increasing the number of compute nodes from 2-8 increases the speedup from 6.4x to 8.2x.

TABLE 4.2 Wall time (w.t.) and relative wall time of water cluster single point gradient calculations using the FMO/RI-MP2 and FMO/MP2 methods.

Water cluster	**#Compute Nodes**	***w.t.* (s) FMO/MP2**	**Aux. Basis**	***w.t.* (s) FMO/RIMP2**	**Speedup**
$(H_2O)_{64}$	2	781.6	cc-pVDZ-RI	121.2	6.4
			aug-cc-pVDZ-RI	146.9	5.3
			cc-pVTZ-RI	148.8	5.3
			aug-cc-pVTZ-RI	158.8	4.9
	4	419.6	cc-pVDZ-RI	61.8	6.8
			aug-cc-pVDZ-RI	71.1	5.9
			cc-pVTZ-RI	87.6	4.8
			aug-cc-pVTZ-RI	96.6	4.3
	8	220.0	cc-pVDZ-RI	26.9	8.2
			aug-cc-pVDZ-RI	29.5	7.5
			cc-pVTZ-RI	35.3	6.2
			aug-cc-pVTZ-RI	34.9	6.3
$(H_2O)_{128}$	2	2791.4	cc-pVDZ-RI	632.9	4.4
			aug-cc-pVDZ-RI	657.1	4.2
			cc-pVTZ-RI	687.3	4.1
			aug-cc-pVTZ-RI	718.5	3.9
	4	2038.3	cc-pVDZ-RI	328.1	6.2
			aug-cc-pVDZ-RI	288.8	7.1
			cc-pVTZ-RI	314.6	6.5
			aug-cc-pVTZ-RI	327.5	6.2
	8	1026.1	cc-pVDZ-RI	122.2	8.4
			aug-cc-pVDZ-RI	133.6	7.7
			cc-pVTZ-RI	143.4	6.2
			aug-cc-pVTZ-RI	152.2	6.7

4.6.4 Multiple node scaling

As demonstrated for the FMO/RI-MP2 energy implementation,[22] the hybrid distributed/shared memory GDDI/OpenMP model can both enhance the computational efficiency and preserve the node linear scaling feature of the FMO framework. In this section, to illustrate the GDDI/OpenMP FMO/RI-MP2 gradient scalability across multiple nodes, the single point analytic gradient of large water clusters containing 1120 and 2615 molecules is calculated using 384-768 KNL nodes. The node scaling is examined by plotting the relative wall time of FMO/RI-MP2 gradient calculations against the number of nodes used. The relative wall time ΔW in an *N*-node calculation is defined as the ratio between the wall time of a 384-node calculation and the wall time of the N-node gradient calculation scaled by a factor of 384. FIGURE 4.3 shows that the relative wall time can be fit to a linear regression equations (101) and (102) with the slope of 0.9736 for gradient calculations on the 1120-water cluster, and a slope of 1.0007 for the cluster of 2615 water molecules. Both slopes are close to the ideal 1.000 linear scaling.

$$\Delta W_{w1120} = 0.9736 \times NODES + 6.8599 \tag{101}$$

$$\Delta W_{w2165} = 1.0007 \times NNODES - 1.2648 \tag{102}$$

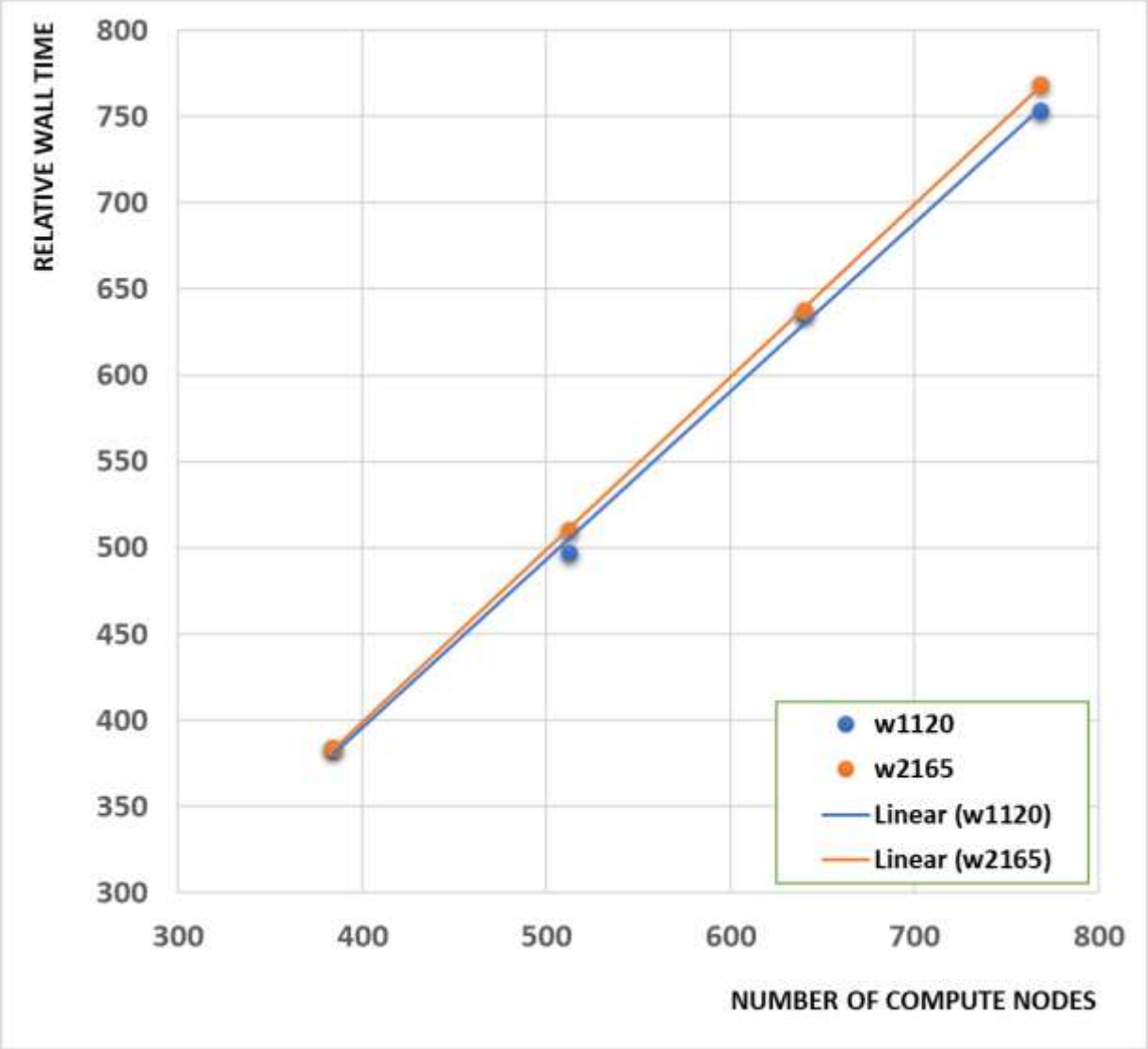


FIGURE 4.3 Relative wall time of FMO/RI-MP2 gradient calculations for clusters of 1120 and 2165 water molecules.

4.7 Concluding remarks

The FMO/RI-MP2 analytic gradient has been derived and implemented in GAMESS using the hybrid distributed/shared memory GDDI/OpenMP model.[22,23] The FMO/RI-MP2 gradient consists of separate internal fragment, ESP and response terms. The RI approximation has been applied to evaluate the internal fragment gradient term whose output is the MP2 density matrix and

internal fragment Lagrangian. The application of the RI approximation and the hybrid parallel model can fully preserve the accuracy of full FMO/MP2 analytic gradient, speed up calculations by a factor of 3.9-8.0x, and maintain the node scalability of the FMO framework. Specifically, benchmark calculations for clusters that contain 16-64 water molecules using the 6-31G(d,p) AO basis set and the auxiliary basis sets cc-pVDZ-RI and aug-cc-pVTZ-RI show that the correlation energy error is below 1.0 kcal/mol and the gradient error is in the range of $10^{-5} - 10^{-7}$ (Hartree/Bohr). The FMO/RI-MP2 energy and gradient are also adequate for MD simulations: the energy is conserved in a NVE MD simulation for water cluster of that contains 16 molecules. Finally, the GDDI/OpenMP FMO/RI-MP2 gradient implementation preserves the scalability of the FMO framework in large scale calculations (e.g., in single point gradient calculations for water cluster of 1120 and 2615 molecules using 384-768 64-core KNL nodes).

Acknowledgements. This work was supported by a grant from the Department of Energy Exascale Computing Project (ECP), administered by the Ames Laboratory. The authors gratefully acknowledge the Argonne Leadership Computing Facility (ALCF) at the Argonne National Laboratory (ANL) and the National Energy Research Scientific Computing Center (NERSC) for providing CPU time and technical support.

References

(1) Kitaura, K.; Sawai, T.; Asada, T.; Nakano, T.; Uebayasi, M. Pair Interaction Molecular Orbital Method: An Approximate Computational Method for Molecular Interactions. *Chem. Phys. Lett.* 1999, *312*, 319–324.

(2) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment Molecular Orbital Method: An Approximate Computational Method for Large Molecules. *Chem. Phys. Lett.* 1999, *313*, 701–706.

(3) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment

Molecular Orbital Method: Application to Polypeptides. *Chem. Phys. Lett.* 2000, *318*, 614–618.

(4) Nakano, T.; Kaminuma, T.; Sato, T.; Fukuzawa, K.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment Molecular Orbital Method: Use of Approximate Electrostatic Potential. *Chem. Phys. Lett.* 2002, *351*, 475–480.

(5) Nagata, T.; Fedorov, D. G.; Kitaura, K. Mathematical Formulation of the Fragment Molecular Orbital Method BT - Linear-Scaling Techniques in Computational Chemistry and Physics: Methods and Applications; Zalesny, R., Papadopoulos, M. G., Mezey, P. G., Leszczynski, J., Eds.; Springer Netherlands: Dordrecht, 2011; pp 17–64.

(6) Gordon, M. S.; Schmidt, M. W. Chapter 41 - Advances in Electronic Structure Theory: GAMESS a Decade Later.; Frenking, G., Kim, K. S., Scuseria, G. E. B. T.-T. and A. of C. C., Eds.; Elsevier: Amsterdam, 2005; pp 1167–1189.

(7) Fedorov, D. G.; Olson, R. M.; Kitaura, K.; Gordon, M. S.; Koseki, S. A New Hierarchical Parallelization Scheme: Generalized Distributed Data Interface (GDDI), and an Application to the Fragment Molecular Orbital Method (FMO). *J. Comput. Chem.* 2004, *25*, 872–880.

(8) Nagata, T.; Fedorov, D. G.; Kitaura, K. Derivatives of the Approximated Electrostatic Potentials in the Fragment Molecular Orbital Method. *Chem. Phys. Lett.* 2009, *475*, 124–131.

(9) Steinmann, C.; Fedorov, D. G.; Jensen, J. H. Effective Fragment Molecular Orbital Method: A Merger of the Effective Fragment Potential and Fragment Molecular Orbital Methods. *J. Phys. Chem. A* 2010, *114*, 8705–8712.

(10) Bertoni, C.; Gordon, M. S. Analytic Gradients for the Effective Fragment Molecular Orbital Method. *J. Chem. Theory Comput.* 2016, *12*, 4743–4767.

(11) 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.

(12) Aquilante, F.; Delcey, M. G.; Pedersen, T. B.; Fdez. Galván, I.; Lindh, R. Inner Projection Techniques for the Low-Cost Handling of Two-Electron Integrals in Quantum Chemistry. *Mol. Phys.* 2017, *115*, 2052–2064.

(13) Røeggen, I.; Wisløff-Nilssen, E. On the Beebe-Linderberg Two-Electron Integral Approximation. *Chem. Phys. Lett.* 1986, *132*, 154–160.

(14) Røeggen, I.; Johansen, T. Cholesky Decomposition of the Two-Electron Integral Matrix in Electronic Structure Calculations. *J. Chem. Phys.* 2008, *128*, 194107.

(15) Koch, H.; Sánchez de Merás, A.; Pedersen, T. B. Reduced Scaling in Electronic Structure Calculations Using Cholesky Decompositions. *J. Chem. Phys.* 2003, *118*, 9481–9484.

(16) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* 1973, *58*, 4496–4501.

(17) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral Approximations for LCAO-SCF

Calculations. *Chem. Phys. Lett.* 1993, *213*, 514–518.

(18) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: Optimized Auxiliary Basis Sets and Demonstration of Efficiency. *Chem. Phys. Lett.* 1998, *294*, 143–152.

(19) Okiyama, Y.; Nakano, T.; Yamashita, K.; Mochizuki, Y.; Taguchi, N.; Tanaka, S. Acceleration of Fragment Molecular Orbital Calculations with Cholesky Decomposition Approach. *Chem. Phys. Lett.* 2010, *490*, 84–89.

(20) Ishikawa, T.; Kuwata, K. Fragment Molecular Orbital Calculation Using the RI-MP2 Method. *Chem. Phys. Lett.* 2009, *474*, 195–198.

(21) Ishikawa, T.; Kuwata, K. RI-MP2 Gradient Calculation of Large Molecules Using the Fragment Molecular Orbital Method. *J. Phys. Chem. Lett.* 2012, *3*, 375–379.

(22) Pham, B. Q.; Gordon, M. S. A Hybrid Distributed/Shared Memory Model For the RI-MP2 Method in the Fragment Molecular Orbital Framework. *(submitted)*.

(23) Mironov, V.; Alexeev, Y.; Fedorov, D. G. Multithreaded Parallelization of the Energy and Analytic Gradient in the Fragment Molecular Orbital Method. *Int. J. Quantum Chem.* 2019, *119*, e25937.

(24) Handy, N. C.; Schaefer, H. F. On the Evaluation of Analytic Energy Derivatives for Correlated Wave Functions. *J. Chem. Phys.* 1984, *81*, 5031–5033.

(25) Nagata, T. (永田武史); Brorsen, K.; Fedorov, D. G.; Kitaura, K. (北浦和夫); Gordon, M. S. (強首領真空). Fully Analytic Energy Gradient in the Fragment Molecular Orbital Method. *J. Chem. Phys.* 2011, *134*, 124115.

(26) Nagata, T.; Fedorov, D. G.; Ishimura, K.; Kitaura, K. Analytic Energy Gradient for Second-Order Møller-Plesset Perturbation Theory Based on the Fragment Molecular Orbital Method. *J. Chem. Phys.* 2011, *135*, 44110.

(27) Nagata, T.; Fedorov, D. G.; Kitaura, K. Importance of the Hybrid Orbital Operator Derivative Term for the Energy Gradient in the Fragment Molecular Orbital Method. *Chem. Phys. Lett.* 2010, *492*, 302–308.

(28) Fedorov, D. G.; Jensen, J. H.; Deka, R. C.; Kitaura, K. Covalent Bond Fragmentation Suitable To Describe Solids in the Fragment Molecular Orbital Method. *J. Phys. Chem. A* 2008, *112*, 11808–11816.

(29) Nishimoto, Y.; Fedorov, D. G. Adaptive Frozen Orbital Treatment for the Fragment Molecular Orbital Method Combined with Density-Functional Tight-Binding. *J. Chem. Phys.* 2018, *148*, 64115.

(30) Fedorov, D. G.; Avramov, P. V.; Jensen, J. H.; Kitaura, K. Analytic Gradient for the Adaptive Frozen Orbital Bond Detachment in the Fragment Molecular Orbital Method. *Chem. Phys. Lett.* 2009, *477*, 169–175.

(31) Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. A Derivation of the Frozen-Orbital Unrestricted Open-Shell and Restricted Closed-Shell Second-Order Perturbation Theory Analytic Gradient Expressions. *Theor. Chem. Acc.* 2003,

110, 233–253.

(32) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. The Distributed Data Interface in GAMESS. *Comput. Phys. Commun.* 2000, *128*, 190–200.

(33) Theta Computer at ALFC. https://www.alcf.anl.gov/theta.

(34) Cori Computer at NERSC https://www.nersc.gov/users/computational-systems/cori/.

(35) Brorsen, K. R.; Minezawa, N.; Xu, F.; Windus, T. L.; Gordon, M. S. Fragment Molecular Orbital Molecular Dynamics with the Fully Analytic Energy Gradient. *J. Chem. Theory Comput.* 2012, *8*, 5008–5012.

CHAPTER 5. CAN ORBITALS REALLY BE OBSERVED IN STM EXPERIMENTS?

A Viewpoint published in The Journal of Physical Chemistry A

Buu Q. Pham and Mark S. Gordon

The scanning-tunneling microscopy (STM) technique has become a fundamental tool to probe surfaces by employing the quantum tunneling effect. Unfortunately, several papers have (incorrectly) claimed to have used novel STM techniques to observe specific molecular orbitals (MOs), in contradiction to the seminal contributions of Werner Heisenberg and Max Born to the fundamentals of quantum mechanics. In this Viewpoint, a brief analysis of the MO concept serves as a reminder that orbitals are simply mathematical constructs that are introduced into quantum mechanics to provide a route to an approximate solution of the Schrodinger equation. Orbitals, therefore, are not observables, claims to the contrary notwithstanding.

A fundamental concept in quantum mechanics is the Heisenberg Uncertainty Principle.[1] For the purpose of this note, the main point is that a wave function, *any* wave function, cannot be thought of as a trajectory or as an observable in the same way that it is possible to consider a trajectory in classical mechanics. Rather, one can speak only of probabilities of finding particles (e.g., electrons) in particular volume elements. In view of the Heisenberg Uncertainty Principle, an important fundamental postulate (the Born postulate)[2] of quantum mechanics is that the probability of finding an electron in a volume element dV is $\Psi^*\Psi dV$, where Ψ is a wave function that describes the electron and Ψ^* is the complex conjugate of Ψ. The wave function amplitude $\Psi^*\Psi$ is interpreted as the probability density. All observable atomic or molecular properties are

determined by the probability and a corresponding quantum mechanical operator, *not* by the wave function itself. Wave functions, even *exact* wave functions are not observables.

The exact wave function is not obtainable for any but the simplest systems, such as the particle in the box or the harmonic oscillator. To obtain an approximate wave function for an atomic or molecular species, that is, an approximate solution to the electronic Schrodinger equation (within the Born-Oppenheimer approximation), one often introduces the concept of one-electron functions (i.e., orbitals) such that the approximate wave function is taken to be an antisymmetrized product of these orbitals. This so-called "mean field" approximation is embedded in both Hartree-Fock (HF) theory and in the most commonly used implementations of density functional theory (DFT). It is important to stress here that these orbitals (called molecular orbitals in HF theory and Kohn-Sham (KS) orbitals in DFT) are merely mathematical constructs that are used to obtain a route to an approximate solution to the Schrodinger equation. Furthermore, since an arbitrary unitary transformation of a wave function does not alter its amplitude (probability density) or consequently, any observable properties, an MO basis used to build the wave function is not unique. For example, the two very different sets of canonical and localized MOs can form bases to build wave functions of the same density, which represent the same electronic state of the system. Therefore, MOs are not unique and are not observable.

Unfortunately, many authors have claimed to image MOs, particularly the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). In particular, these include claims to have observed the HOMO and LUMO of fullerene (2000),[3] nitrogen (2004),[4] pentacene (2005),[5,6] Co^{II} tetraphenylporphyrin (2008),[7] polyaromatic hydrocarbons,[8] water (2014),[9] and 3,4,9,10-perylene-tetracarboxylic acid dianhydride (2017).[10] STM

experiments seem to be the most common tool used to “probe” frontier orbitals. A recent Perspective[11] in Nature Chemistry claimed that STM images of a molecule chemically/physically absorbed on a substrate are images of (frontier) orbitals that are coupled with the electronic state of the matrix. When the coupling to the matrix is eliminated (e.g., by inserting ultrathin insulating NaCl films), the author claimed that the STM images become the images of “native”[6] (frontier) MOs. Based on the foregoing discussion, this cannot be true. Orbitals are not observables.

The origin of the misinterpretation of STM images as orbitals can be understood as follows. When a bias voltage is applied to a tip positioned very close to a material surface, electron tunneling can occur. The variation of the tunneling current can be visualized as the STM image. The tunneling current at a particular space point is proportional to the density of states (DOS) of the material at that point, also called the Local DOS. The DOS at a particular energy level results from the coupling of the electronic state at that energy level to perturbation sources (e.g., electronic states of adjacent molecules, external fields). The DOS at an energy level of an unperturbed state (e.g., an isolated molecule) is small and proportional to the degeneracy order at that energy level. In the STM technique, a local potential applied to the tip serves as the perturbation; therefore, the DOS is non-zero. In many molecular systems, a high DOS occurs at about the Fermi level of energy, which is usually between the first ionization and electron affinity energies. These energy levels can be *roughly approximated* by the HOMO and LUMO energies. Unfortunately, several authors have made the incorrect leap to conclude that they are actually observing the HOMO and LUMO themselves.

Claims regarding the observation of orbitals have been made even earlier than those mentioned above. In 1999, Zuo *et* al.[12] attributed an X-ray image to be a *d*-orbital of copper in Cu_2O. This paper received some much-needed attention from, for example, Scerri,[13] and Mulder[14], both of whom explained in some detail that orbitals are *not* observable. The papers by Scerri and by Mulder are important and should be read carefully by those who assert that orbitals can be observed in experiments. Unfortunately, these two papers have been ignored by several authors.

To summarize, *i*) any property of a system is only fully represented by the total density; *ii*) orbitals are simply mathematical constructs used to build the (approximate) wave function and then the density; *iii*) orbitals can therefore not be associated directly with an observable molecular property; *iv*) Orbitals are non-unique, since the energy is invariant to any unitary transformation among the (HF or KS) orbitals within a given subspace (e.g., doubly occupied space).

Acknowledgements. The authors thank Professor James Evans and Dr. Da Jiang Liu for helpful discussions, and the US Air Force Office of Scientific Research (FA9550-14-1-0306) for support.

References

(1) Heisenberg, W. Über Den Anschaulichen Inhalt Der Quantentheoretischen Kinematik Und Mechanik. *Zeitschrift für Phys.* **1927**, *43*, 172–198.

(2) Born, M. Statistical Interpretation of Quantum Mechanics. *Science (80-.).* **1955**, *122*, 675 LP – 679.

(3) Pascual, J. I.; Gómez-Herrero, J.; Rogero, C.; Baró, A. M.; Sánchez-Portal, D.; Artacho, E.; Ordejón, P.; Soler, J. M. Seeing Molecular Orbitals. *Chem. Phys. Lett.* **2000**, *321*, 78–82.

(4) Itatani, J.; Levesque, J.; Zeidler, D.; Niikura, H.; Pépin, H.; Kieffer, J. C.; Corkum, P. B.; Villeneuve, D. M. Tomographic Imaging of Molecular Orbitals. *Nature* **2004**, *432*, 867–871.

(5) Soe, W.-H.; Manzano, C.; De Sarkar, A.; Chandrasekhar, N.; Joachim, C. Direct Observation of Molecular Orbitals of Pentacene Physisorbed on Au(111) by Scanning Tunneling Microscope. *Phys. Rev. Lett.* **2009**, *102*, 176102.

(6) Repp, J.; Meyer, G.; Stojković, S. M.; Gourdon, A.; Joachim, C. Molecules on Insulating Films: Scanning-Tunneling Microscopy Imaging of Individual Molecular Orbitals. *Phys. Rev. Lett.* **2005**, *94*, 26803.

(7) Weber-Bargioni, A.; Auwärter, W.; Klappenberger, F.; Reichert, J.; Lefrançois, S.; Strunskus, T.; Wöll, C.; Schiffrin, A.; Pennec, Y.; Barth, J. V. Visualizing the Frontier Orbitals of a Conformationally Adapted Metalloporphyrin. *ChemPhysChem* **2008**, *9*, 89–94.

(8) Villagomez, C. J.; Zambelli, T.; Gauthier, S.; Gourdon, A.; Stojkovic, S.; Joachim, C. STM Images of a Large Organic Molecule Adsorbed on a Bare Metal Substrate or on a Thin Insulating Layer: Visualization of HOMO and LUMO. *Surf. Sci.* **2009**, *603*, 1526–1532.

(9) Guo, J.; Meng, X.; Chen, J.; Peng, J.; Sheng, J.; Li, X.-Z.; Xu, L.; Shi, J.-R.; Wang, E.; Jiang, Y. Real-Space Imaging of Interfacial Water with Submolecular Resolution. *Nat. Mater.* **2014**, *13*, 184.

(10) Puschnig, P.; Boese, A. D.; Willenbockel, M.; Meyer, M.; Lüftner, D.; Reinisch, E. M.; Ules, T.; Koller, G.; Soubatch, S.; Ramsey, M. G.; et al. Energy Ordering of Molecular Orbitals. *J. Phys. Chem. Lett.* **2017**, *8*, 208–213.

(11) Gross, L. Recent Advances in Submolecular Resolution with Scanning Probe Microscopy. *Nat. Chem.* **2011**, *3*, 273.

(12) Zuo, J. M.; Kim, M.; O'Keeffe, M.; Spence, J. C. H. Direct Observation of D-Orbital Holes and Cu–Cu Bonding in Cu2O. *Nature* **1999**, *401*, 49–52.

(13) Scerri, E. R. Have Orbitals Really Been Observed? *J. Chem. Educ.* **2000**, *77*, 1492.

(14) Mulder, P. Are Orbitals Observable? *Hyle* **2011**, *17*.

CHAPTER 6. THERMODYNAMICS AND KINETICS OF GRAPHENE CHEMISTRY: A GRAPHENE HYDROGENATION PROTOTYPE STUDY

A paper published in Physical Chemistry Chemical Physics

Buu Q. Pham and Mark S. Gordon

Abstract

The thermodynamic and kinetic controls of graphene chemistry are studied computationally using a graphene hydrogenation reaction and polyaromatic hydrocarbons to represent the graphene surface. Hydrogen atoms are concertedly chemisorped onto the surface of graphene models of different shapes (i.e., all-zigzag, all-armchair, zigzag-armchair mixed edges) and sizes (i.e., from 16-42 carbon atoms). The *second*-order *Z*-averaged perturbation theory (ZAPT2) method combined with Pople double and triple zeta basis sets are used for all calculations. It is found that both the net enthalpy change and the barrier height of graphene hydrogenation at graphene edges are lower than at their interior surfaces. While the thermodynamic product distribution is mainly determined by the remaining π-islands of functionalized graphenes (*PCCP* **2013**, *15*, 3725-35), the kinetics of the reaction is primarily correlated with the localization of the electrostatic potential of the graphene surface.

6.1 Introduction

Graphenes are “chickenwire-like” carbon sheets, which have been studied intensively and extensively during the last few decades. The first successful effort to isolate graphene was to peel off carbon layers from highly oriented pyrolytic graphite (HOPG).[1] The limited amount of graphene laboriously obtained using this mechanical exfoliation method was a turning point for a new generation of carbon-based materials.[2] Since then, free standing,[3] or deposited[4] forms of graphene have proved to be powerful materials with a number of novel features,[5] suggesting numerous potential applications.[6] Nevertheless, *two*-dimensional (2D) graphene has limitations that potentially restrict its applicability. For example, the lack of a significant band gap in graphenes[1] limits their use in on-off electronic devices. Therefore, a wide range of chemical engineering and functionalization methods have been used to manipulate graphene energy band gaps. For instance, by taking advantage of the quantum confinement effect, “gapless” 2D graphene sheets can chemically be transformed into 1D graphene ribbons,[7] or finite-size graphene clusters (quantum dots)[8] with desired electronic band gaps. The graphene band gap can also be tuned by functionalizing graphene edges,[9] doping the graphene surface with non-metal elements,[10] or converting graphene to graphane [hydrogenated graphene].[11] Understanding and controlling graphene chemistry, therefore, plays an important role in graphene technology.[12]

Except for 2D sheets, all other forms of graphene are expected to have physical and chemical properties that vary with respect to their shapes and sizes. For instance, by increasing the length of linear polyacenes, considered the simplest series of graphene ribbons, their electronic ground states have been reported to evolve from a closed-shell singlet state (e.g., benzene) to an open-

shell triplet state (e.g., octa- and nona-acene).[13] The trend of decreasing energy differences between closed-shell ground states and open-shell excited states has been reported in a number of studies.[14,15] The potential energy surfaces of graphene excited states can involve complex processes with the ground state and with each other, such as avoided crossings, conical intersections, and internal conversions.[16] Shape and size effects are, therefore, prominent in studies of graphene cluster chemistry.

An important question regarding graphene chemistry is the location of its active centers, which control the distribution of functionalized products. An active center can be defined in terms of thermodynamics, which determines the positions of the most stable products; or in terms of kinetics, which determines where functionalized products will form at faster rates. Most (theoretical and experimental) graphene chemistry studies have mainly focused on the thermodynamics of reactions,[17,18] with little attention paid to the kinetics, even though understanding the kinetics of graphene chemistry may provide an important tool in graphene technology. For instance, a recent advanced graphene etching technique[19] is based on the possibility that a graphene edge can be oxidized faster than its interior surface. Although a detailed mechanism for such a process is not known, the exploitation of approximate kinetic information has led to large benefits.

Therefore, this work presents a preliminary study of the thermodynamics and kinetics (barrier heights) of atom addition to graphene. Graphene hydrogenation by hydrogen atoms is used as a prototype for this purpose.

6.2 Computational methods

To account for shape and size effects, different graphene cluster models are used for the current investigation. Figure 6.1 depicts *three* series of graphene models. The first series (Figure 6.1a, b and c) are all-zigzag edged graphene models ***ZN***, which can be formed by adding phenanthrene-exterior patterns to a pyrene molecule. The index ***N*** is the number of linearly fused benzene rings in the middle row of the graphene model. The other graphene models include those that have an all-armchair edge (***AA***, Figure 6.1d), and a combined zigzag-armchair mixed edge (***ZA***, Figure 6.1e).

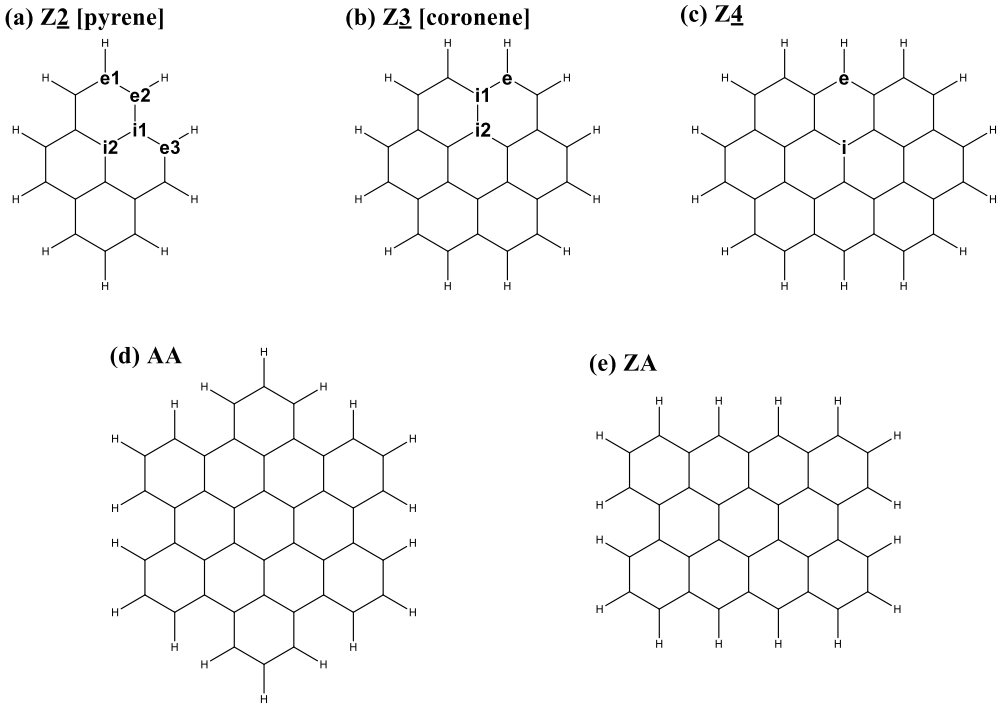


Figure 6.1 *a-c) all-zigzag (**ZN**), d) all-armchair (**AA**), and e) zigzag-armchair (**ZA**) mixed edged graphene models. The notations **e** and **i** indicate an edge site and an interior site, respectively.*

A hydrogen atom can occupy *three* possible positions on a graphene surface: a bridge [**B**] site (on the middle of a CC bond), a hollow [**H**] site (on the center of a benzene ring), or an atop [**A**] site (on the top of a carbon atom). A hydrogen atom can only form a CH covalent bond with the graphene surface at an **A** site, while the **B** and **H** adsorptions give rise to saddle points, or unbound states.[20,21] The current investigation studies the relative energies and barrier heights of CH bond formation on the surface of graphene models.

All stationary points along the reaction coordinates for each CH bond formation have been located. Minima and first-order saddle points were confirmed by calculating and diagonalizing the Hessian matrices. The minimum energy paths connecting the transition state with desired reactants and products have also been calculated. All calculated energy profiles of the hydrogenation reactions are corrected for vibrational zero-point energies (ZPE).

All calculations were done using second-order Møller-Plesset perturbation theory (MP2) combined with the Pople double (6-31G(d)) and triple-zeta (6-311+G(d,p)) split valence shell basis sets. To avoid spin contamination, for open-shell systems the second-order Z-averaged perturbation theory method (ZAPT2)[22,23] is used. All calculations were done using GAMESS.[24] Calculated results are parsed and visualized using the MacMolplt visualization package.[25]

6.3 Results and discussion

6.3.1 Computational model calibration

Quantum investigations of graphene chemistry are computationally expensive, and the computational cost rapidly increases as the graphene size increases. Therefore, selecting an appropriate theoretical model that compromises between the accuracy and the cost of

calculations is important. Since an essential part of this investigation is the *relative* chemical reactivity among active centers on graphene surfaces, it is worthwhile to determine if a relatively small basis set can provide useful qualitative data compared to a larger basis set. This could be important as the size of the graphene cluster increases. Therefore, this section compares the energy profiles of graphene hydrogenation using different basis sets. The pyrene molecule is used to represent the graphene surface. At absolute zero, the energy profiles are represented by the reaction enthalpy $\left(\Delta H_{\varpi}^{0}\right)$ and the reaction barrier $\left(\Delta H_{\varpi}^{\ddagger}\right)$, which are obtained from ZPE-corrected electronic energies of reactants $\left(E_{\varpi,R}^{zpe}\right)$, transition state $\left(E_{\varpi,T}^{zpe}\right)$ and products $\left(E_{\varpi,P}^{zpe}\right)$ in Equations (1) and (2). The subscripts R, P and T stand for reactants, products and transition state, respectively, and the subscript ϖ denotes for the reaction center; i.e., **e1**-**3**, **i1-2**. The results of this analysis are summarized in Table 6.1.

$$\Delta H_{\varpi}^{0} = E_{\varpi,P}^{zpe} - E_{\varpi,R}^{zpe} \tag{1}$$

$$\Delta H_{\varpi}^{\ddagger} = E_{\varpi,T}^{zpe} - E_{\varpi,R}^{zpe} \tag{2}$$

Table 6.1 *MP2 heats* $\left(\Delta H_{\varpi}^{0}\right)$ *and barriers* $\left(\Delta H_{\varpi}^{\ddagger}\right)$ *in kcal/mol of the pyrene hydrogenation reaction at different carbon atoms* (cf. Figure 6.1a).

Methods	6-31G(d)		6-311G(d)		6-311+G(d, p)	
Sites (ϖ)	**ΔH^0**	**$\Delta H^{\ddagger}$**	**ΔH^0**	**$\Delta H^{\ddagger}$**	**ΔH^0**	**$\Delta H^{\ddagger}$**
e1	-2.5	18.3	-3.3	17.2	-7.9	13.7
e2	-19.8	15.8	-20.3	14.5	-25.2	11.5
e3	-13.0	14.9	-16.8	13.9	-21.9	11.1
i1	10.7	21.8	8.6	19.7	3.9	15.8
i2	7.0	20.5	5.1	17.8	0.2	13.9

While pyrene (**Z2**, Figure 6.1a) is a relatively small graphene model, it has both edge (Figure 6.1: **e1**, **e2** and **e3**) and interior (**i1** and **i2**) carbon atoms, thereby providing a good starting point to represent a graphene surface. Full energy profiles of pyrene hydrogenation at all symmetrically unique carbon atoms were evaluated using the ZAPT2 method combined with the Pople double-(6-31G(d)), triple-zeta (6-311G(d)) and triple zeta plus diffuse (6-311+G(d,p)) basis sets.

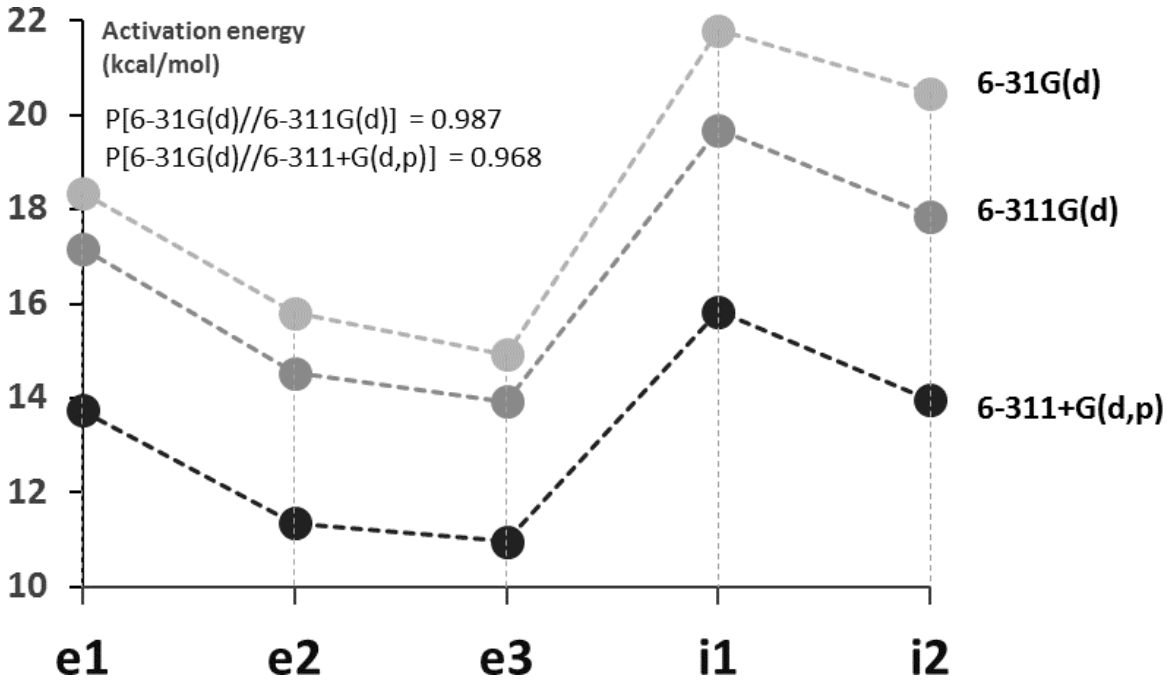


Figure 6.2 *Hydrogenation barriers calculated using three basis sets. P is a Pearson correlation coefficient.*

Table 6.1 shows that when the size of the basis set increases, the reaction becomes more exothermic and the barrier height decreases. For example, increasing the basis set from double (6-31G(d)) to triple (6-311G(d)) zeta quality, the reaction barrier goes down by 6.2-11.4%. Increasing the basis set further to 6-311+G(d,p) reduces the barrier height further by 25.1-32.2%. For instance, the hydrogenation barrier at **i1** monotonically decreases from 20.5-13.9 kcal/mol

when increasing from double to triple zeta basis sets. This barrier using the restricted open shell singles and doubles coupled cluster (ROCCSD) and the corresponding method with perturbative triples [ROCCSD(T)], with the cc-pVDZ basis set is in the range of 10.1-11.3 kcal/mol.[21] Therefore, increasing the basis set brings MP2 barriers closer to the couple cluster results.

Interestingly, the energy profiles obtained from small and large basis sets are highly correlated. As shown in Figure 6.2, the pyrene hydrogenation barriers obtained using different basis sets are almost parallel to each other. Quantitatively, the Pearson coefficient (P),[26] which varies from 0.000 (no correlation) to 1.000 (perfect correlation) can be used to evaluate the correlation among data sets. For two *N*-point data sets $\{x_i\}_N$ and $\{y_i\}_N$, whose means are *X* and *Y*, the Pearson coefficient can be formulated in Eq. (3).

$$P\left[\{x_i\}_N, \{y_i\}_N\right] = \frac{\left|\sum_i^N (x_i - X)(y_i - Y)\right|}{\left(\sum_i^N (x_i - X)^2 \sum_i^N (y_i - Y)^2\right)^{1/2}} \quad (3)$$

Figure 6.2 shows that the Pearson coefficient between barriers calculated using the 6-31G(d) and the 6-311G(d) basis sets is very close to 1.000; i.e., P[6-31G(d)//6-311G(d)]=0.987. Similarly, P[6-31G(d)//6-311G(d)]=0.968. Therefore, it appears to be reasonable to use the ZAPT2/6-31G(d) method to investigate the *relative* chemical reactivities of graphene surfaces.

6.3.2 Thermodynamics of graphene hydrogenation

Graphene can be hydrogenated when exposed to an atomized hydrogen plasma in vacuum,[11] or in a molecular hydrogen atmosphere at high temperature and pressure.[27] Graphene

hydrogenation by hydrogen atoms is closely related to a number of studies in astrophysics, where pyrene and extended polyaromatic hydrocarbons (PAH) [considered to be graphene clusters] are known to catalyze the formation of a molecular hydrogen atmosphere on dust grain surfaces.[28] The process happens in a stepwise reaction, whose the first step is the chemisorption of hydrogen atoms on the PAH surface. This is followed by clustering-recombining,[29] and/or H-abstraction[30] to form hydrogen molecules. It is notable that an excess of adsorbed hydrogen atoms is found to accumulate on the edges of an extended PAH.[31,32] Graphene edges are, thus, expected to be more reactive than their inner surfaces, at least from a thermodynamic point of view. The results shown in Table 6.1 and Figure 6.2 are consistent with this observation. In this section, the relative thermodynamic energies of carbon atoms at different positions on graphene surfaces are elucidated in terms of the heats of hydrogenation at the ZAPT2/6-31G(d) level of theory.

Table 6.1 shows that hydrogenation at a pyrene edge site (**e1**, **e2** or **e3**: See Figure 6.1) is more exothermic than at an interior site (**i1** or **i2**). The variation of the heat of hydrogenation (at 0K) is:

$$\Delta H_{\varpi}^{0}\left(kcal/mol\right):e2\ \left(-19.8\right)\ <\ e3\ \left(-13.0\right)\ <\ e1\ \left(-2.5\right)\ <\ i2\ \left(7.0\right)\ <\ i1\ \left(10.7\right) \quad (4)$$

So, the heat of hydrogenation at the edge is thermodynamically much more favorable than at the interior regions. Similarly, the heat of hydrogenation at the edge of the coronene surface (Figure 6.1b) is distinctly more negative than at its interior surface:

$$\Delta H_{\varpi}^{0}\left(kcal/mol\right):e\ \left(-5.7\right)\ <\ i1\ \left(3.1\right)\ <\ i2\ \left(4.1\right) \quad (5)$$

The difference in the chemical affinity toward a hydrogen atom between a carbon at the graphene edge and a carbon on the inner surface can be attributed in part to the relative stabilities of the remaining π-islands of the hydrogenated products. The graphene edge may be

more reactive since the graphene π-conjugating system seems to be less disrupted when the hydrogenation occurs at this peripheral region. Qualitatively, the relative stability of π-islands can be evaluated using several familiar indices; e.g., the Kekulé index (the maximum number of Kekulé benzene structures obtained from the arrangement of π-bonds), the resonance energy evaluated using the Huckel method,[33] or the nuclear independent chemical shift (NICS) index.[34]

More quantitatively, the hydrogenation reaction may be thought of as breaking a double bond in a pool of other double bonds on the graphene surface. Therefore, the relative stabilities of different regions of the graphene surface can be evaluated by comparing the stabilities of these double bonds with a standard, isolated double bond (e.g., of a ethylene molecule), using the concept of an isodesmic reaction first introduced by Pople and co-workers.[35,36] In an isodesmic reaction, the number of electron pairs and bonds of each type is conserved; consequently, energy errors due to the limitation of the computational model used are minimized. The relative stabilities of double bonds on the graphene surface is evaluated using the isodesmic reaction depicted in Figure 6.3, in which a graphene model is hydrogenated by an ethyl radical instead of a hydrogen atom to form an ethylene molecule and graphene hydrogenated product. The double bond on the surface of the graphene model is transformed into the double bond in the ethylene molecule, which allows a direct comparison of their relative stabilities. The heats (kcal/mol) of these isodesmic reactions for ***Z2***, ***Z3*** and ***Z4*** are depicted in Figure 6.3a, b, and c, respectively. For pyrene (***Z2***) and coronene (***Z3***), the entire cluster is examined; for the larger model ***Z4***, to save computation time, only the edge and interior surface carbon sites that result in hydrogenated products with Cs symmetry are considered.

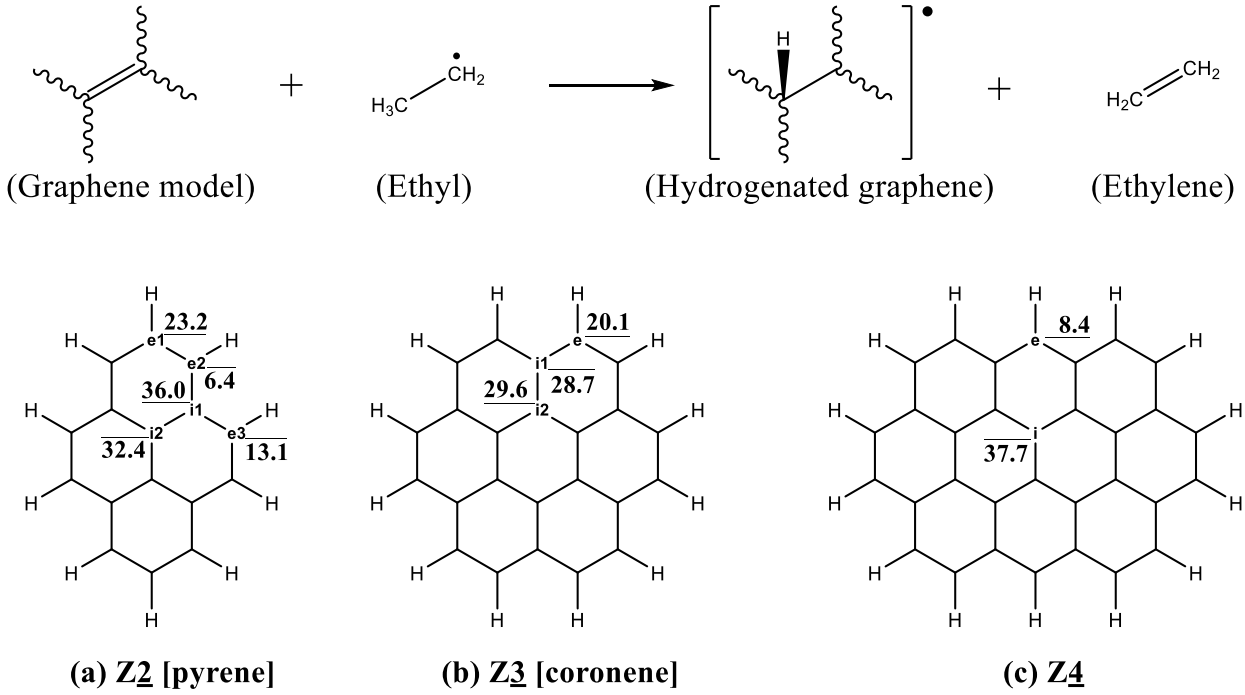


Figure 6.3 *Heat of graphene isodesmic hydrogenation reaction using ethyl radical.*

The heats of the isodesmic hydrogenation reactions (Figure 6.3) for ***Z2*** to ***Z4*** are all positive. This indicates that the double bonds in the graphene models used in this investigation are more stable than the isolated double bond in the ethylene molecule. For instance, the heat of isodesmic hydrogenation varies from 6.4 kcal/mol (the edge carbon **e2** of ***Z2***, Figure 6.3a) up to 37.7 kcal/mol (the interior carbon ***i*** of ***Z4***, Figure 6.3c). The heat of isodesmic hydrogenation also reveals that it costs much less energy to break double bonds at graphene edges than those located in the graphene inner surface, indicating that the graphene edges are more reactive than their interior surfaces.

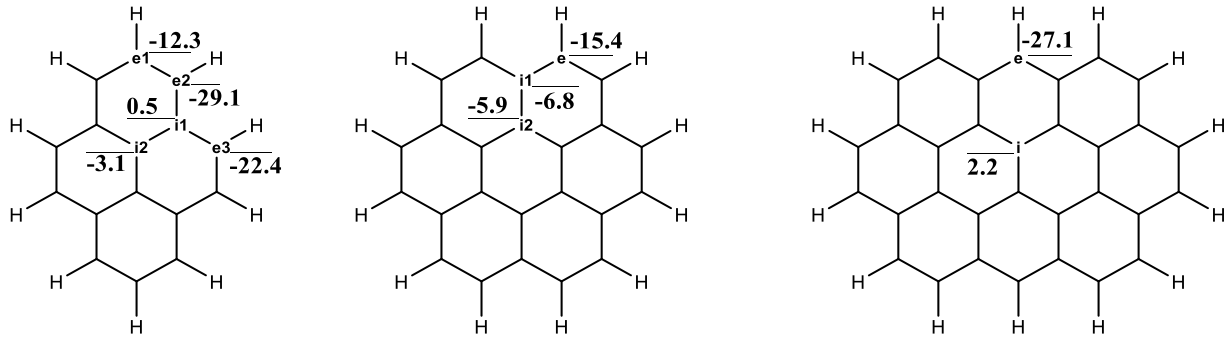
If one combines the isodesmic reaction discussed above with the experimental bond dissociation of ethyl radical,[37] as indicated in Eqs. (6) and (7), additional interesting energetic (Eq. (8)) comparisons are revealed. The results depicted in Figure 6.4 show that graphene

hydrogenation at the edges are much more exothermic than at the interior surfaces, suggesting that graphene hydrogenation at the edges is feasible, and thermodynamically more favorable than at the interior surfaces.

$$R+\left(C_2H_5\right)^{\bullet}\rightarrow\left(RH\right)^{\bullet}+C_2H_4 \quad , \quad \Delta H^{isodesmic} \tag{6}$$

$$\left(C_2H_5\right)^{\bullet}\rightarrow H^{\bullet}+C_2H_4 \quad , \quad \Delta H^{dissociation}=35.5\left(kcal/mol\right) \tag{7}$$

$$R+H^{\bullet}\rightarrow\left(RH\right)^{\bullet} \quad , \quad \Delta H=\Delta H^{isodesmic}-\Delta H^{dissociation} \tag{8}$$



(a) Z2 [pyrene] **(b) Z3 [coronene]** **(c) Z4**

Figure 6.4 *Heat of graphene hydrogenation* (*Eq.* (8)) *obtained by combining isodesmic reaction* (*Eq.* (6)) and ethyl radical dissociation (Eq. (7)).

6.3.3 Kinetics of graphene hydrogenation

Note that the addition of a hydrogen atom to the double bond of ethylene is known to be almost barrierless (e.g., ca. 3.0 kcal/mol).[38] While an experimental value for the graphene hydrogenation barrier is not available, the barrier height for graphite hydrogenation has been found to be larger than that of ethylene (e.g., ca. 5.0 kcal/mol).[39] In comparison with graphenes, the interlayer interaction in graphite probably activates the π-system. The heat of the isodesmic

graphene hydrogenation reaction (cf. Figure 6.3) demonstrates that the π-system in graphene is much more stable than the isolated double bond in ethylene. Therefore, the barrier heights in graphene hydrogenation are expected to be larger than that of graphite and ethylene. These observations are consistent with the calculated barrier for the pyrene (*Figure 6.5*), and coronene (Figure 6.6) graphene cluster models discussed below.

The reaction barriers for graphene hydrogenation also divide the graphene surfaces into edge regions and interior regions. For example, at the ZAPT2/6-31G(d) level of theory, the barriers (at 0K) for pyrene hydrogenation are:

$$\Delta H_{\varpi}^{\ddagger}\left(kcal/mol\right): e3\ \left(14.9\right) < e2\ \left(15.8\right) < e1\ \left(18.3\right) < i2\ \left(20.5\right) < i1\ \left(21.8\right) \qquad (9)$$

While the thermodynamic product distribution of graphene chemistry is controlled by its products,[18] the current understanding of the kinetic indicators of graphene chemistry is limited.

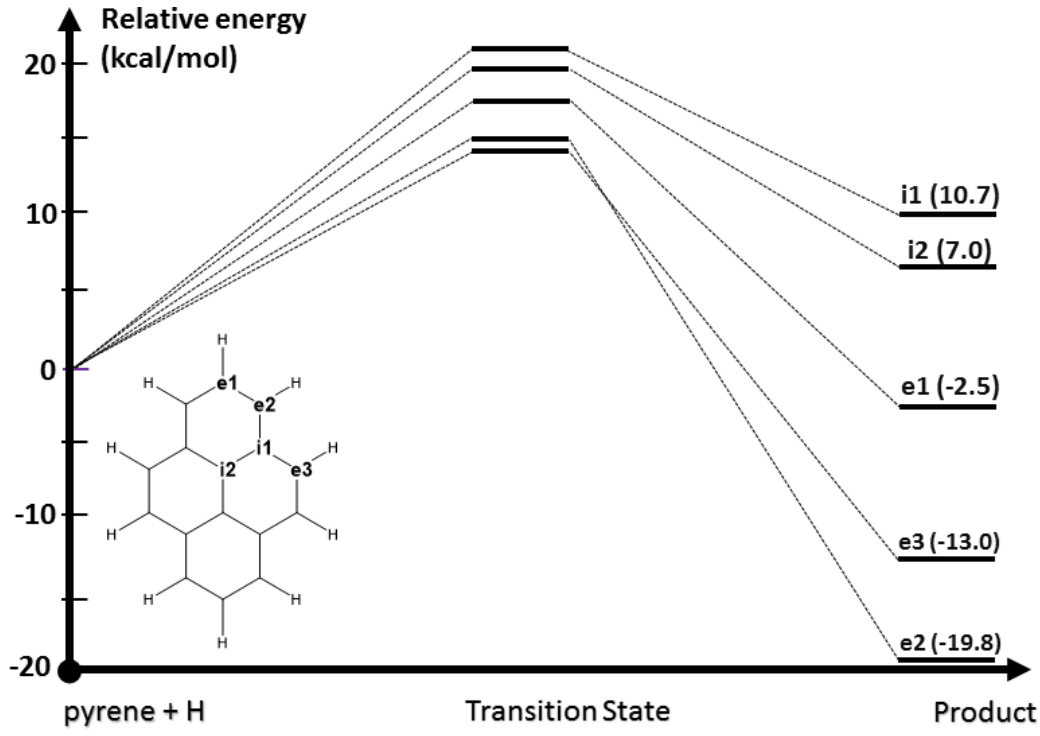
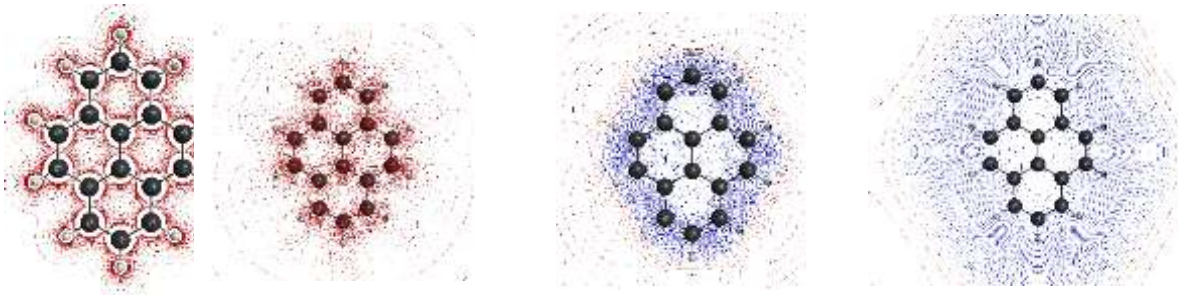


Figure 6.5 *ZAPT2/6-31G(d) energy profiles of pyrene hydrogenation*

To examine the reaction kinetics, consider the molecular electrostatic potential (MEP),[40] which maps the response of the molecular potential to a test charge. A MEP map, therefore, can provide insight into the kinetics of chemical processes by identifying regions of nucleophilicity and electrophilicity. For planar graphenes, it is convenient to examine 2D MEP maps generated on planes at different distances from their surfaces. Figure 6.6 depicts pyrene MEP maps evaluated in planes from 0-3Å above the pyrene molecular plane. On planes close to the pyrene molecular surface (0-1 Å), the positive potential [red contours] is dominant and evenly distributed among all carbon sites. However, at distances 2Å and above, which is the area of valence interaction, the negative potential [blue contours] becomes prominent and is localized at the pyrene edge, while the inner surface shows little response. Since one can think of the approaching H atom to be electrophilic, the MEP maps are consistent with the observed barrier heights. The pyrene MEPs are in agreement with the relative chemical reactivity of pyrene edge, and the experimental evidence of graphene edge states.[41,42]



a) 0 Å ***b) 1 Å*** ***c) 2 Å*** ***d) 3 Å***

Figure 6.6 *MP2/6-31G(d) 2D MEPs generated in planes from 0-3 Å above the pyrene molecular plane. Positive, and negative potentials are in red, and blue, respectively.*

Similarly, the reaction barriers for coronene hydrogenation are smaller at the edge carbons than at the interior carbons, as illustrated in *Figure 6.7*a. The 2D MEP map generated on the

plane 2Å above the coronene molecular surface (Figure 6.7b) once again demonstrates that the reactivity is localized on the edges.

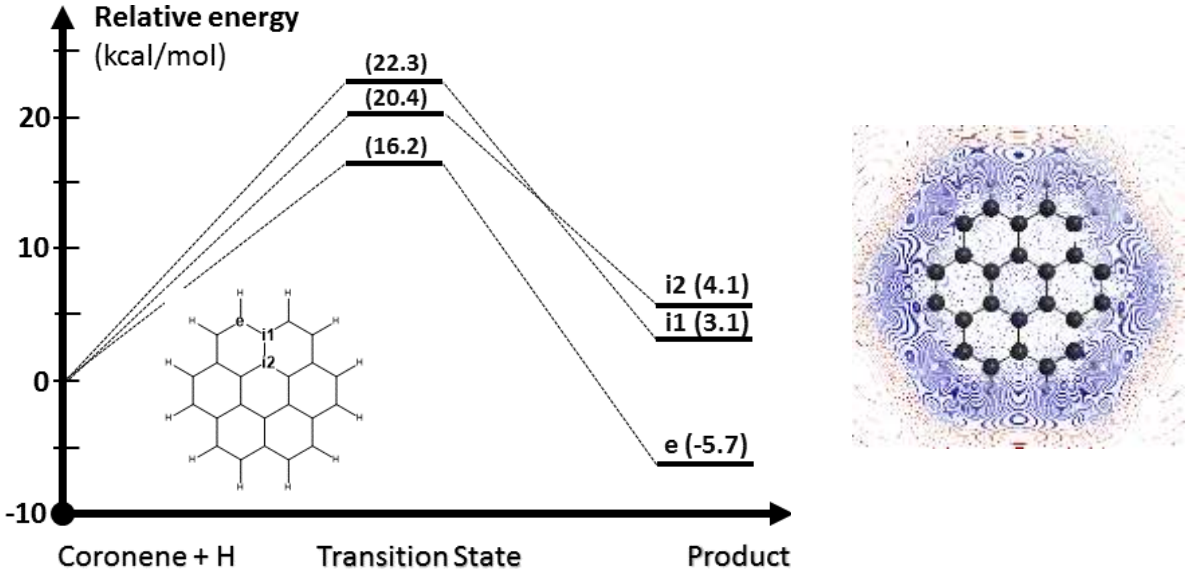
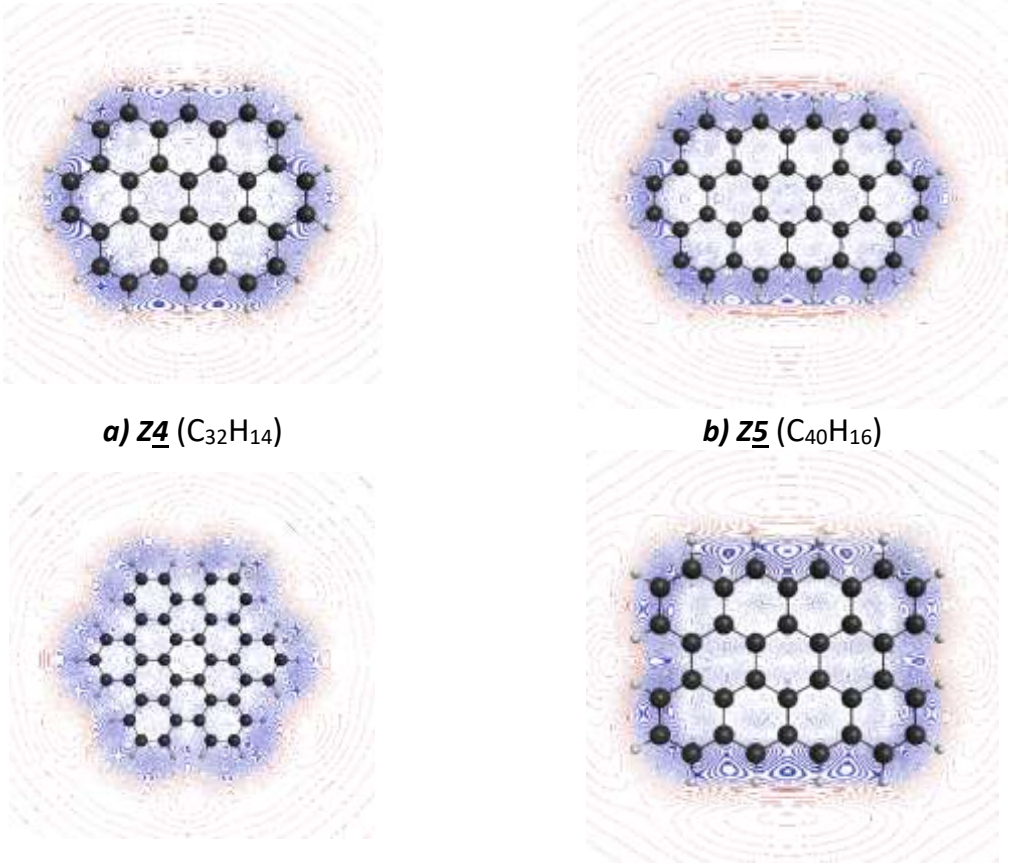


Figure 6.7 *(a) The ZAPT2/6-31G(d) energy profiles of coronene hydrogenation; and (b) the MP2/6-31G(d) 2D MEP map generated on a plane 2Å above the coronene molecular plane. Positive, and negative potentials are in red, and blue, respectively.*

The 2D MEPs generated on planes 2Å above the ***Z4*** (32 carbon atoms) and ***Z5*** (40 carbon atoms) models are depicted in Figure 6.8a and 6.7b, respectively. Both MEPs have high negative contours concentrated on their edges, while the interior region has little response. Other topologies are also illustrated, including the all-armchair (***AA***, Figure 6.1c), and zigzag-armchair mixing (***ZA***, Figure 6.1d) edge graphene models. All of their 2D MEPs shown in Figure 6.8c-d show the same features as those discussed above, again suggesting the susceptibility of the edge carbon atoms to hydrogen atom attack.



a) Z4 ($C_{32}H_{14}$) ***b) Z5*** ($C_{40}H_{16}$)

c) AA ($C_{42}H_{18}$) ***d) ZA*** ($C_{36}H_{16}$)

Figure 6.8 *2D MEPs generated on planes 2 Å above the molecular planes of a)* ***Z4*** *($C_{32}H_{14}$), b)* ***Z5*** *($C_{40}H_{16}$), c)* ***AA*** *($C_{42}H_{18}$) and d)* ***ZA*** *($C_{36}H_{16}$) graphene models. Positive and negative potentials are in red, and blue, respectively.*

6.4 Concluding remarks

Graphene hydrogenation reactions have been explored using second order Moller-Plesset perturbation calculations. Graphene edges are found to be both thermodynamically and kinetically more reactive than the interior surface. The thermodynamics and kinetics of graphene are controlled in part by different factors. While the thermodynamic product distribution is guided by the nature of the π–charge delocalization, the reaction kinetics are governed by

localization of the molecular electrostatic potential in the valence region above (and below) the rings that kinetically prefer attack at the graphene edge.

Acknowledgements. This work was supported by a Department of Defense HPCMP Applications Software Initiative (HASI) grant.

References

(1) Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Zhang, Y.; Dubonos, S. V.; Grigorieva, I. V.; Firsov, A. A. Electric Field in Atomically Thin Carbon Films. *Science (80-.).* **2004**, *306*, 666–669.

(2) Geim, A. K.; Novoselov, K. S. The Rise of Graphene. *Nat. Mater.* **2007**, *6*, 183–191.

(3) Eberlein, T.; Bangert, U.; Nair, R. R.; Jones, R.; Gass, M.; Bleloch, A. L.; Novoselov, K. S.; Geim, A.; Briddon, P. R. Plasmon Spectroscopy of Free-Standing Graphene Films. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2008**, *77*, 233406.

(4) Report, C. Mind the Gap ! Mind the Gap ! *Nat. Mater.* **2012**, *96*, 10–12.

(5) Castro Neto, A. H.; Guinea, F.; Peres, N. M. R.; Novoselov, K. S.; Geim, A. K. The Electronic Properties of Graphene. *Rev. Mod. Phys.* **2009**, *81*, 109–162.

(6) Novoselov K. S.; Fal'ko V. I.; Colombo L.; Gellert P. R.; Schwab M. G.; Kim K. A Roadmap for Graphene. *Nature* **2012**, *490*, 192–200.

(7) Han, M. Y.; Özyilmaz, B.; Zhang, Y.; Kim, P. Energy Band-Gap Engineering of Graphene Nanoribbons. *Phys. Rev. Lett.* **2007**, *98*, 206805.

(8) Jin, Z.; Owour, P.; Lei, S.; Ge, L. Graphene, Graphene Quantum Dots and Their Applications in Optoelectronics. *Curr. Opin. Colloid Interface Sci.* **2015**, *20*, 439–453.

(9) Wagner, P.; Ewels, C. P.; Adjizian, J. J.; Magaud, L.; Pochet, P.; Roche, S.; Lopez-Bezanilla, A.; Ivanovskaya, V. V.; Yaya, A.; Rayson, M.; et al. Band Gap Engineering via Edge-Functionalization of Graphene Nanoribbons. *J. Phys. Chem. C* **2013**, *117*, 26790–26796.

(10) Geng, D.; Yang, S.; Zhang, Y.; Yang, J.; Liu, J.; Li, R.; Sham, T. K.; Sun, X.; Ye, S.; Knights, S.

Nitrogen Doping Effects on the Structure of Graphene. *Appl. Surf. Sci.* **2011**, *257*, 9193–9198.

(11) Elias, D. C.; Nair, R. R.; Mohiuddin, T. M. G.; Morozov, S. V.; Blake, P.; Halsall, M. P.; Ferrari, A. C.; Boukhvalov, D. W.; Katsnelson, M. I.; Geim, A. K.; et al. Control of Graphene's Properties by Reversible Hydrogenation: Evidence for Graphane. *Science (80-.).* **2009**, *323*, 610–613.

(12) Georgakilas, V.; Otyepka, M.; Bourlinos, A. B.; Chandra, V.; Kim, N.; Kemp, K. C.; Hobza, P.; Zboril, R.; Kim, K. S. Functionalization of Graphene: Covalent and Non-Covalent Approaches, Derivatives and Applications. *Chem. Rev.* **2012**, *112*, 6156–6214.

(13) Tönshoff, C.; Bettinger, H. F. Photogeneration of Octacene and Nonacene. *Angew. Chemie - Int. Ed.* **2010**, *49*, 4125–4128.

(14) Pham, B. Q.; Truong, T. N. Electronic Spin Transitions in Finite-Size Graphene. *Chem. Phys. Lett.* **2012**, *535*, 75–79.

(15) Ramakrishna Matte, H. S. S.; Subrahmanyam, K. S.; Rao, C. N. R. Novel Magnetic Properties of Graphene: Presence of Both Ferromagnetic and Antiferromagnetic Features and Other Aspects. *J. Phys. Chem. C* **2009**, *113*, 9982–9985.

(16) Pham, B. Q.; Nguyen, V. H.; Truong, T. N. Size Dependence of Graphene Chemistry: A Computational Study on CO Desorption Reaction. *Carbon N. Y.* **2016**, *101*, 16–21.

(17) Liu, L.; Wang, L.; Gao, J.; Zhao, J.; Gao, X.; Chen, Z. Amorphous Structural Models for Graphene Oxides. *Carbon N. Y.* **2012**, *50*, 1690–1698.

(18) Page, A. J.; Chou, C. P.; Pham, B. Q.; Witek, H. A.; Irle, S.; Morokuma, K. Quantum Chemical Investigation of Epoxide and Ether Groups in Graphene Oxide and Their Vibrational Spectra. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3725–3735.

(19) Wang, X.; Dai, H. Etching and Narrowing of Graphene from the Edges. *Nat. Chem.* **2010**, *2*, 661–665.

(20) Jeloaica, L.; Sidis, V. DFT Investigation of the Adsorption of Atomic Hydrogen on a Cluster-Model Graphite Surface. *Chem. Phys. Lett.* **1999**, *300*, 157–162.

(21) Wang, Y.; Qian, H. J.; Morokuma, K.; Irle, S. Coupled Cluster and Density Functional Theory Calculations of Atomic Hydrogen Chemisorption on Pyrene and Coronene as Model Systems for Graphene Hydrogenation. *J. Phys. Chem. A* **2012**, *116*, 7154–7160.

(22) Fletcher, G. D.; Gordon, M. S.; Bell, R. S. Gradient of the ZAPT2 Energy. *Theor. Chem. Acc.* **2002**, *107*, 57–70.

(23) Aikens, C. M.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. Scalable Implementation of Analytic Gradients for Second-Order Z-Averaged Perturbation Theory Using the Distributed Data Interface. *J. Chem. Phys.* **2006**, *124*, 14107.

(24) W., S. M.; K., B. K.; A., B. J.; T., E. S.; S., G. M.; H., J. J.; Shiro, K.; Nikita, M.; A., N. K.; Shujun, S.; et al. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **2004**, *14*, 1347–1363.

(25) Bode, B. M.; Gordon, M. S. MacMolPlt: A Graphical User Interface for GAMESS. *J. Mol. Graph. Model.* **1998**, *16*, 133–138.

(26) M Mukaka, M. Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* **2012**, *24*, 69–71.

(27) Smith, D.; Howie, R. T.; Crowe, I. F.; Simionescu, C. L.; Muryn, C.; Vishnyakov, V.; Novoselov, K. S.; Kim, Y. J.; Halsall, M. P.; Gregoryanz, E.; et al. Hydrogenation of Graphene by Reaction at High Pressure and High Temperature. *ACS Nano* **2015**, *9*, 8279–8283.

(28) Puget, J. L.; Léger, A. A New Component of the Interstellar Matter: Small Grains and Large Aromatic Molecules. *Annu. Rev. Astron. Astrophys.* **2003**, *27*, 161–198.

(29) Hornekær, L.; Rauls, E.; Xu, W.; Šljivančanin, Ž.; Otero, R.; Stensgaard, I.; Lægsgaard, E.; Hammer, B.; Besenbacher, F. Clustering of Chemisorbed H(D) Atoms on the Graphite (0001) Surface Due to Preferential Sticking. *Phys. Rev. Lett.* **2006**, *97*, 186102.

(30) Zecho, T.; Güttler, A.; Sha, X.; Lemoine, D.; Jackson, B.; Küppers, J. Abstraction of D Chemisorbed on Graphite (0001) with Gaseous H Atoms. *Chem. Phys. Lett.* **2002**, *366*, 188–195.

(31) Mennella, V.; Hornekær, L.; Thrower, J.; Accolla, M. The Catalytic Role of Coronene for Molecular Hydrogen Formation. *Astrophys. J. Lett.* **2012**, *745*, L2.

(32) Katz, N.; Furman, I.; Biham, O.; Pirronello, V.; Vidali, G. Molecular Hydrogen Formation on Astrophysically Relevant Surfaces. *Astrophys. J.* **1999**, *522*, 305–312.

(33) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39*, 1397–1412.

(34) Moran, D.; Stahl, F.; Bettinger, H. F.; Schaefer, H. F.; Schleyer, P. v. R. Towards Graphite: Magnetic Properties of Large Polybenzenoid Hydrocarbons. *J. Am. Chem. Soc.* **2003**, *125*, 6746–6752.

(35) Benson, S. W. III - Bond Energies. *J. Chem. Educ.* **1965**, *42*, 502.

(36) Hehre, W. J.; Ditchfield, R.; Radom, L.; Pople, J. A. Molecular Orbital Theory of the Electronic Structure of Organic Compounds. V. Molecular Theory of Bond Separation. *J. Am. Chem. Soc.* **1970**, *92*, 4796–4801.

(37) Hase, W. L.; Schlegel, H. B. Resolution of a Paradox Concerning the Forward and Reverse Rate Constants for Ethyl .Dblarw. Atomic Hydrogen + Ethylene. *J. Phys. Chem.* **1982**, *86*, 3901–3904.

(38) Hase, W. L.; Schlegel, H. B.; Balbyshev, V.; Page, M. An Ab Initio Study of the Transition State and Forward and Reverse Rate Constants for C2H5 ⇌ H + C2H4. *J. Phys. Chem.* **1996**, *100*, 5354–5361.

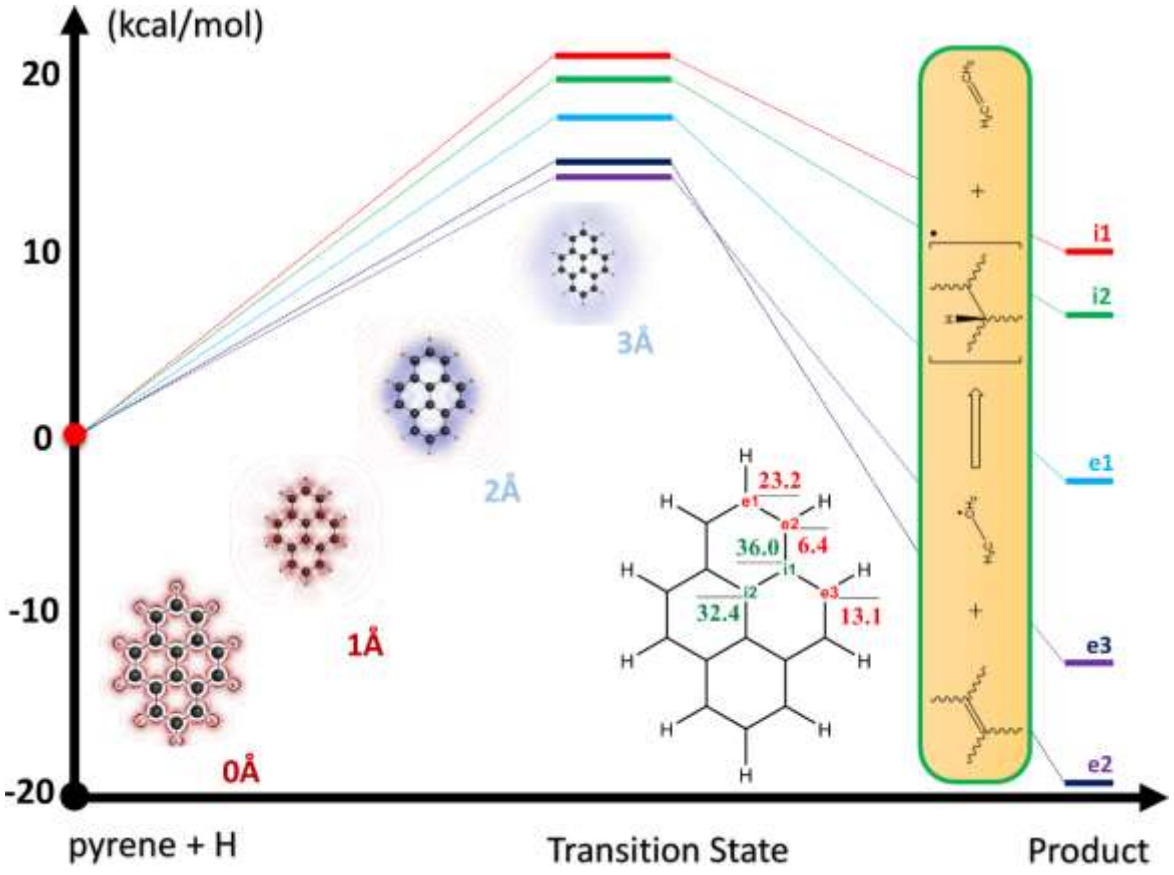
(39) Aréou, E.; Cartry, G.; Layet, J.-M.; Angot, T. Hydrogen-Graphite Interaction: Experimental Evidences of an Adsorption Barrier. *J. Chem. Phys.* **2011**, *134*, 14701.

(40) Politzer, P.; Truhlar, D. G. Introduction: The Role of the Electrostatic Potential in Chemistry BT - Chemical Applications of Atomic and Molecular Electrostatic Potentials: Reactivity, Structure, Scattering, and Energetics of Organic, Inorganic, and Biological Systems; Politzer, P., Truhlar, D. G., Eds.; Springer US: Boston, MA, 1981; pp 1–6.

(41) Kobayashi, Y.; Fukui, K.; Enoki, T.; Kusakabe, K.; Kaburagi, Y. Observation of Zigzag and Armchair Edges of Graphite Using Scanning Tunneling Microscopy and Spectroscopy. *Phys. Rev. B* **2005**, *71*, 193406.

(42) Niimi, Y.; Matsui, T.; Kambara, H.; Tagami, K.; Tsukada, M.; Fukuyama, H. Scanning Tunneling Microscopy and Spectroscopy of the Electronic Local Density of States of Graphite Surfaces near Monoatomic Step Edges. *Phys. Rev. B* **2006**, *73*, 85421.

Graphical abstract



CHAPTER 7. SUMMARY AND CONCLUSION

The computational demand, including the number of floating-point operations and the memory storage of *ab initio* calculations, for macromolecular system is huge. Parallel implementations to make use of large computing resources in these calculations are difficult due to memory limitations and the serial parts of the code. Studies in this dissertation suggested a template that enable *ab initio* calculations, particularly the second-order Moller-Plesset perturbation theory, for large systems by combining integral compressors, fragmentation methods, and efficient parallel models. Integral compressors are used to address the bottleneck of the 2-electron integral transformation (from the atomic orbital to the molecular orbital basis) as well as the memory demand of correlated *ab initio* methods. The fragmentation methods partition a system into small pieces that helps to remove a large part of the redundant integrals (e.g., short-rang exchange integrals), avoid large matrix processing (e.g., diagonalizing Fock matrix of the whole large molecule). In addition, fragmentation methods enable 2-level parallelism; e.g., the first level distributes compute resources among fragments, the second level provides parallel code within fragment calculations. The multilayer parallelization basically removes serial parts in a fragmentation code while retaining the computational scalability when using a large number of compute nodes. Finally, an efficient parallel model is used enhance the performance. For instance, by using a hybrid distributed/shared memory model, the communication overhead and the memory footprint are significantly reduced. The light-weight threads with natural resource sharing character can efficiently make use of new computer

hardware, especially multicore processors. The template of hybrid-fragmentation-compressor has initially been built and presented in chapters 2-4 as follows.

Chapter 2 reviews popular integral compressors including the resolution-of-the-identity (RI) approximation, Cholesky decomposition, stochastic RI approximation, and the natural auxiliary function method. A new compression scheme for the 4-2ERI matrix has also been formulated by combining the RI approximation with the singular value decomposition (SVD). The SVD-RI method introduces a rigorous accuracy controller yielding a precise systematic error that fully preserves the accuracy of relevant relative properties (e.g., potential energy surface, binding energy on the MP2 correlation energy test calculations).

Chapter 3 illustrates the combination of the RI approximation, the fragment molecular orbital (FMO) method, and the hybrid distributed/shared memory GDDI/OpenMP model for the MP2 correlation energy method. This results in a speedup of a factor of 10x relative to the available GDDI FMO/RI-MP2 code in calculations on medium to large molecular systems (e.g., of 417-11,259 atoms). The hybrid parallel model can also preserve the linear scaling characteristics of the FMO framework. Calculations for water clusters that contain 2,165 molecules (11,259 atoms) exhibits linear node scaling (e.g., when the number of 64-core compute nodes is varied from 256-768).

Chapter 4 applies the hybrid-fragmentation-compressor template to the MP2 analytic gradient. Similar to the energy, the application of the RI approximation and the hybrid parallel model to the gradient can fully preserve the accuracy of the FMO/MP2 analytic gradient while speeding up calculations by a factor of 3.9-8.0x, and maintaining the node scalability of the FMO

framework. The test calculations were carried out on systems of up to 11,259 atoms and using up to 768 64-core compute nodes.

For future developments, the SVD-RI subroutine will be optimized to further reduce the computational cost. More study to develop the analytic gradient is necessary to bring the approximation to practical applications. More variants of the hybrid-fragmentation-compressor template can be developed for optimal combinations. For instance, the integral compressor can be chosen to be the Cholesky decomposition, or even pure SVD for the 4-2ERI; the fragmentation method can be selected from the large fragmentation pool e.g., EFP, QM/EFP, FMO, EFMO; other efficient computational models can be applied to enhance the performance of the code implementation (e.g., heavy computation can be processed with accelerators).