An Improved Algorithm for Analytical Gradient Evaluation in Resolution-of-the-Identity Second-Order Møller-Plesset Perturbation Theory: Application to Alanine Tetrapeptide Conformational Analysis

ROBERT A. DISTASIO, JR., RYAN P. STEELE, YOUNG MIN RHEE, YIHAN SHAO, MARTIN HEAD-GORDON
Department of Chemistry, University of California, Berkeley, California 94720

Received 30 March 2006; Revised 29 June 2006; Accepted 20 July 2006
DOI 10.1002/jcc.20604
Published online 11 January 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We present a new algorithm for analytical gradient evaluation in resolution-of-the-identity second-order Møller-Plesset perturbation theory (RI-MP2) and thoroughly assess its computational performance and chemical accuracy. This algorithm addresses the potential I/O bottlenecks associated with disk-based storage and access of the RI-MP2 *t*-amplitudes by utilizing a semi-direct batching approach and yields computational speed-ups of approximately 2–3 over the best conventional MP2 analytical gradient algorithms. In addition, we attempt to provide a straightforward guide to performing reliable and cost-efficient geometry optimizations at the RI-MP2 level of theory. By computing relative atomization energies for the G3/99 set and optimizing a test set of 136 equilibrium molecular structures, we demonstrate that satisfactory relative accuracy and significant computational savings can be obtained using Pople-style atomic orbital basis sets with the existing auxiliary basis expansions for RI-MP2 computations. We also show that RI-MP2 geometry optimizations reproduce molecular equilibrium structures with no significant deviations ($>$0.1 pm) from the predictions of conventional MP2 theory. As a chemical application, we computed the extended-globular conformational energy gap in alanine tetrapeptide at the extrapolated RI-MP2/cc-pV(TQ)Z level as 2.884, 4.414, and 4.994 kcal/mol for structures optimized using the HF, DFT (B3LYP), and RI-MP2 methodologies and the cc-pVTZ basis set, respectively. These marked energetic discrepancies originate from differential intramolecular hydrogen bonding present in the globular conformation optimized at these levels of theory and clearly demonstrate the importance of long-range correlation effects in polypeptide conformational analysis.

© 2007 Wiley Periodicals, Inc. J Comput Chem 28: 839–856, 2007

Key words: second-order Møller-Plesset theory; RI-MP2; MP2 analytical gradient; resolution-of-the-identity approximation; alanine; Pople-style basis sets; auxiliary basis sets; density-fitting; equilibrium geometries; force field parameters

Introduction

The ability to accurately predict molecular equilibrium structures has been one of the primary driving forces for the use of computational chemistry. In fact, theoretical predictions of molecular geometries have been deemed so reliable that many researchers consider them a viable alternative to experimental structure determination; at this point, the usefulness of theory is no longer restricted to cases where structural information might be extremely difficult, or even impossible, to obtain using current state-of-the-art experimental methods.[1–4] For many experimentalists, performing geometry optimizations using standard electronic structure methods has become a routine practice during the analysis of experimental findings, and more often than not, even calculations performed on a personal desktop computer can provide valuable insight and even influence future experimental directions.

The recent works of Helgaker and coworkers[5,6] have assessed the performance of the standard hierarchy of *ab initio* models in the structural optimization of molecular systems containing first and second row atoms. This hierarchy of electronic structure methods starts with the mean-field Hartree-Fock (HF) approximation,[7,8] continues with the simplest correlation treatment, second-order Møller-Plesset perturbation theory (MP2),[9,10] and then the higher level coupled-cluster theories that include single and double excitations (CCSD)[11] as well as perturbative triples (CCSD(T)).[12,13] Although these methods are readily available in most computational software packages, their practical use is often heavily restricted by the size of the molecular system of interest. In fact, geometry optimizations of molecular systems comprised of only

*This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

Correspondence to: M. Head-Gordon; e-mail: mhg@bastille.cchem.berkeley.edu

Contract/grant sponsors: NIH SBIR; National Science Foundation

50–200 heavy atoms—systems that are orders of magnitude smaller than those found in computational biochemistry and nanotechnology research—are even outside the practical limits of conventional MP2 theory! Density functional theory (DFT) has therefore emerged as the most computationally effective method for treating electron correlation with moderate accuracy, but at a fraction of the cost associated with MP2 computations.[14–16]

When MP2 theory is formulated in a basis of canonical molecular orbitals, i.e., the orthonormal set of eigenfunctions that diagonalize the Fock matrix, one finds a fifth-order computational dependence on the system size.[17] For comparison, current implementations of the iterative HF and DFT methods can demonstrate nearly linear scaling[18–25] for reasonably extended molecular systems despite a formal fourth-order dependence on system size. This relatively high computational cost, coupled with the need for large atomic orbital basis sets for reliable results,[26] limits MP2 from the regime of large molecular systems. However, MP2 theory does have several advantages over DFT worth mentioning. Unlike DFT, which suffers from the well-documented self-interaction problem[27,28] and the present inability to account for long-range correlation effects,[29] MP2 theory provides a sound physical description of dispersion interactions[30] while accounting for approximately 80–90% of the correlation energy.[2] In fact, optimizations using analytical MP2 gradient theory have furnished equilibrium geometries and molecular properties that are more reliable than HF,[31] popular DFT alternatives (for example, see Table 8-4 of ref. 16 and references therein), and, in some cases, even CCSD.[32] As an electronic structure theory, the main limitation associated with MP2 is its diminished performance when dealing with radicals and/or systems with small HOMO/LUMO energy gaps.[32,33] Additional limitations include overestimation of π-π stacking interactions[34] and poor treatment of systems that contain certain transition metals[35]—drawbacks that can certainly limit the use of MP2 theory in the study of metalloproteins and several classes of enzymatic reactions. However, when dealing with most large closed-shell molecular systems, it still remains that the relatively high computational cost associated with MP2 theory is by far its main limitation.

One of many[36–42] promising approaches to reducing this prohibitively high computational cost involves the use of the resolution-of-the-identity (RI) approximation.[43,44] Also referred to as the density-fitting (DF) approach, this technique uses linear combinations of auxiliary basis set functions to approximate the atomic orbital pair densities present in the MP2 ansatz. In doing so, the incorporation of the RI approximation into MP2 theory (RI-MP2) successfully reduces the computational prefactor associated with MP2 computations but leaves the underlying fifth-order scaling untouched. Alternative approaches often directly attack this fifth-order scaling via the use of models that employ localized molecular orbitals, namely, the classic work of Saebø and Pulay[45] (LMP2), the more recent atomistic truncation schemes of Head-Gordon and coworkers[46–48] (DIM/TRIM MP2), and others.[49,50] By combining the RI approximation with a simultaneous reduction in the MP2 correlation space, as was accomplished by the DF-LMP2[51] and RI-TRIM MP2[52] methodologies, even greater savings can be realized—in these cases, both the underlying scaling and computational prefactor are substantially reduced.

Although LMP2[53] and DF-LMP2[54] analytical gradients are currently available, the underlying localization scheme in both of these methodologies is heavily based on numerical thresholding, and as a result, these methods fail to produce continuous and smooth potential energy surfaces.[55] On the other hand, RI-MP2 analytical gradient evaluation produces continuous and smooth potential energy surfaces, and the theory behind RI-MP2 analytical gradients involving closed-shell systems has been understood for quite some time now.[56] In reexamining this initial report, which contains the first serial RI-MP2 algorithm in the literature to date, we were able to identify some potential I/O bottlenecks in the suggested algorithm. Despite these I/O challenges, however, the current implementation of the RI-MP2 geometry optimization procedure is able to computationally outperform even the most advanced conventional MP2 gradients,[57–61] thereby extending the regime of molecular systems that can be investigated at the MP2 level of theory.

The purpose of this paper is to report a further improvement in RI-MP2 technology by presenting a new algorithm that addresses these potential I/O bottlenecks and achieves higher computational efficiency during RI-MP2 analytical gradient evaluation. This algorithm, which utilizes a semi-direct batching scheme to address the I/O bottlenecks associated with the storage and access of the MP2 t-amplitudes, is presented as an alternative to the direct approach of Hättig and Weigend[62] as part of their RI-CC2 module. With this new algorithm, RI-MP2 theory could then be utilized more effectively for treating large molecules such as polypeptide and polynucleotide model systems relevant for the fundamental parameterization of force fields and other semi-empirical methods. Since the parameters present in most force fields are sometimes derived from the structural predictions of HF or DFT, a comparative analysis of the quality of such parameters seems to be one immediate and powerful application of this new algorithm. Therefore, we present such an analysis in the application section of this work, wherein the structures and energetics associated with extended and globular conformations of alanine tetrapeptide are characterized.

The remainder of this paper is outlined as follows. In the Theory Section, we begin by presenting the working expressions associated with the RI-MP2 analytical gradient in the spin-orbital basis. This general spin-orbital formulation of the RI-MP2 analytical gradient is then specialized to handle both unrestricted open-shell (UHF) and restricted closed-shell (RHF) wave functions within the frozen core (and frozen virtual) approximation. An in-depth discussion of our RI-MP2 gradient algorithm follows in the Algorithm Section, accompanied by a detailed analysis of the associated computational requirements. In the Algorithm Performance Section, we attempt to provide a straightforward guide to the accuracy and computational timings associated with our new RI-MP2 analytical gradient algorithm. The section begins with a short discussion on the use of Pople-style basis sets during RI-MP2 single-point energy evaluations. We then explore the effects of basis set size and convergence/thresholding criteria on the algorithm presented herein. We conclude this section with a concise assessment of the performance of RI-MP2 theory in optimizing 136 test molecules complete with experimentally determined equilibrium bond lengths. In the Chemical Application Section, we offer a chemical application of our RI-MP2 analytical gradient algorithm in which we seek to provide further insight into the parameterization of force fields by carefully characterizing the extended-globular conformational energy gap in alanine tetrapeptide. The paper is then finished with some brief conclusions.

Theory

In this section and the remainder of the paper, indices a,b,c,. . . will refer to canonical virtual molecular orbitals, A,B,C,. . . to frozen virtual molecular orbitals, i,j,k,. . . to canonical occupied molecular orbitals, I,J,K,. . . to frozen occupied molecular orbitals, p,q,r,. . . to general arbitrary molecular orbitals (MO), μ,ν,λ,σ,. . . to atomic orbital basis set functions (AO), and P,Q,R,S,. . . to auxiliary basis set functions (AUX). The act [occ] and vact [virt] limits of summation denote inclusion of active [all] members of the occupied and virtual subspaces, respectively. The inactive subsets of the occupied and virtual subspaces will be denoted by core and vf, respectively.

The RI-MP2 Analytical Gradient in the General Spin-Orbital Basis

The RI-MP2 analytical gradient expression in the general spin-orbital basis is given as[56,63]

$$E^{x}_{\mathrm{RI-MP2}} = 2\sum_{\mu\nu}^{\mathrm{AO}}\sum_{Q}^{\mathrm{AUX}}(\mu\nu|Q)^{x}\Gamma^{Q}_{\mu\nu} - \sum_{RS}^{\mathrm{AUX}}(R|S)^{x}\Gamma^{RS} + \sum_{pq}^{\mathrm{MO}}P^{(2)}_{pq}F^{(x)}_{pq} + \sum_{pq}^{\mathrm{MO}}W^{(2)}_{pq}S^{(x)}_{pq} + \sum_{pq}^{\mathrm{MO}}\overline{W}^{(2)}_{pq}S^{(x)}_{pq} + \sum_{ij}^{\mathrm{occ}}\overline{\overline{W}}^{(2)}_{ij}S^{(x)}_{ij} \qquad (1)$$

In this expression, the first two terms include integral derivatives involving both the AO and AUX basis sets. These terms are unique to RI-MP2 theory and represent the contributions to the analytical gradient via the three- and two-centered forces—these terms directly replace the contraction of the set of four-centered two-electron repulsion integral derivatives (DERIs) with the non-separable correction to the two-particle density matrix (2-PDM) in conventional MP2 theory, i.e., $\sum_{\mu\nu\lambda\sigma}^{\mathrm{AO}}(\mu\nu|\lambda\sigma)^{x}\,\Gamma^{\mathrm{NS}}_{\mu\nu\lambda\sigma}$. These RI-MP2 specific corrections to the 2-PDM are given as

$$\Gamma^{Q}_{\mu\nu} = \sum_{i}^{\mathrm{act}}\sum_{a}^{\mathrm{vact}}C_{\mu i}C_{\nu a}\Gamma^{Q}_{ia} = \sum_{ij}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}}C_{\mu i}C_{\nu a}t^{ab}_{ij}C^{Q}_{jb} \qquad (2)$$

and

$$\Gamma^{\mathrm{RS}} = \sum_{i}^{\mathrm{act}}\sum_{a}^{\mathrm{vact}}C^{R}_{ia}\Gamma^{S}_{ia} = \sum_{ij}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}}C^{R}_{ia}t^{ab}_{ij}C^{S}_{jb} \qquad (3)$$

where $C_{\mu p}$ is an element of the MO coefficient matrix and C^{Q}_{jb} is an element of the RI-MP2 coefficient matrix given by

$$C^{Q}_{jb} = \sum_{P}^{\mathrm{AUX}}(jb|P)(P|Q)^{-1} \qquad (4)$$

The set of first-order amplitudes, t^{ab}_{ij}, are given as the solution to the following set of linear equations:

$$\sum_{k}^{\mathrm{act}}[F_{ik}t^{ab}_{kj} + F_{kj}t^{ab}_{ik}] - \sum_{c}^{\mathrm{vact}}[F_{ac}t^{cb}_{ij} + F_{cb}t^{ac}_{ij}] = (ia\|jb)_{RI} \qquad (5)$$

where $(ia \parallel jb)_{RI}$ are the anti-symmeterized, four-centered, two-electron repulsion integrals (ERIs) in Mulliken notation, $(ia \parallel jb) = (ia \mid jb) - (ib \mid ja)$, evaluated within the RI approximation. In a basis of canonical MOs, these amplitudes take on the following simpler and more familiar form, namely, $t^{ab}_{ij} = (ia \parallel jb)/(\Delta^{ab}_{ij}) = (ia \parallel jb)/(\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b)$, since both the Fock and overlap matrices are diagonal in this representation, and $F_{pq} = \varepsilon_{pq} = \delta_{pq}\varepsilon_p$.

The remaining terms in eq. (1) include contractions of skeleton derivatives (i.e., AO-DERIs that do not include expansion coefficient derivatives and are denoted by superscript parenthetic notation: $^{(x)}$) of Fock ($F^{(x)}_{pq}$) and overlap ($S^{(x)}_{pq}$) matrix elements with various RI-MP2 corrections to the one-particle density matrix (1-PDM). After some algebra, the RI-MP2 corrections to the 1-PDM are given by (undefined blocks in the list below contain only null elements):

The RI-MP2 Corrections to the 1-PDM (P)

Occupied–Occupied Block (o–o)

$$P^{(2)}_{iK} = \sum_{a}^{\mathrm{vact}}\sum_{Q}^{\mathrm{AUX}}\frac{(Ka|Q)\Gamma^{Q}_{ia}}{(\varepsilon_i - \varepsilon_K)} = \frac{1}{2}\sum_{ab}^{\mathrm{vact}}\sum_{j}^{\mathrm{act}}\frac{(Ka\|jb)t^{ab}_{ij}}{(\varepsilon_i - \varepsilon_K)} = P^{(2)}_{Ki} \quad \forall\ i \in \mathrm{act},\ K \in \mathrm{core} \qquad (6)$$

$$P^{(2)}_{ik} = -\frac{1}{2}\sum_{j}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}}t^{ab}_{ij}t^{ab}_{kj} = P^{(2)}_{ki} \quad \forall\ i,k\ \in \mathrm{act} \qquad (7)$$

Virtual–Virtual Block (v–v)

$$P^{(2)}_{ac} = \frac{1}{2}\sum_{ij}^{\mathrm{act}}\sum_{b}^{\mathrm{vact}}t^{ab}_{ij}t^{cb}_{ij} = P^{(2)}_{ca} \quad \forall\ a{,}c \in \mathrm{vact} \qquad (8)$$

$$P^{(2)}_{aC} = \sum_{i}^{\mathrm{act}}\sum_{Q}^{\mathrm{AUX}}\frac{(iC|Q)\Gamma^{Q}_{ia}}{(\varepsilon_a - \varepsilon_C)} = \frac{1}{2}\sum_{ij}^{\mathrm{act}}\sum_{b}^{\mathrm{vact}}\frac{(iC\|jb)t^{ab}_{ij}}{(\varepsilon_a - \varepsilon_C)} = P^{(2)}_{Ca} \quad \forall\ a \in \mathrm{vact},\ C \in \mathrm{vf} \qquad (9)$$

Occupied–Virtual Block (o–v)

$$P^{(2)}_{ai} \equiv Z_{ai} \equiv P^{(2)}_{ia} \quad \forall\ i \in \mathrm{occ},\ a \in \mathrm{virt} \qquad (10)$$

The RI-MP2 Corrections to the Energy-Weighted 1-PDM (W)

Type I

Occupied–Occupied Block (o–o)

$$W^{(2)}_{ik} = -\sum_{a}^{\mathrm{vact}}\sum_{Q}^{\mathrm{AUX}}(ka|Q)\Gamma^{Q}_{ia} = -\frac{1}{2}\sum_{j}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}}(ka\|jb)t^{ab}_{ij} = W^{(2)}_{ki} \quad \forall\ i \in \mathrm{act},\ k \in \mathrm{occ} \qquad (11)$$

Virtual–Virtual Block (v–v)

$$W^{(2)}_{ac} = -\sum_{i}^{\mathrm{act}}\sum_{Q}^{\mathrm{AUX}}(ic|Q)\Gamma^{Q}_{ia} = -\frac{1}{2}\sum_{ij}^{\mathrm{act}}\sum_{b}^{\mathrm{vact}}(ic\|jb)t^{ab}_{ij} = W^{(2)}_{ca} \quad \forall\ a \in \mathrm{vact},\ c \in \mathrm{virt} \qquad (12)$$

Occupied–Virtual Block (o–v)

$$W^{(2)}_{aj} = -\sum_{i}^{\text{act}}\sum_{Q}^{\text{AUX}}(ij|Q)\Gamma^{Q}_{ia} = -\frac{1}{2}\sum_{ik}^{\text{act}}\sum_{b}^{\text{vact}}(ij\|kb)t^{ab}_{ik} = W^{(2)}_{ia} \quad \forall\ a \in \text{vact},\ i \in \text{occ} \quad (13)$$

Type II
Occupied–Occupied Block (o–o)

$$\overline{W}^{(2)}_{ik} = -\frac{1}{2}P^{(2)}_{ik}(\varepsilon_i + \varepsilon_k) = \overline{W}^{(2)}_{ki} \quad \forall\ i \in \text{act},\ k \in \text{occ} \quad (14)$$

Virtual–Virtual Block (v–v)

$$\overline{W}^{(2)}_{ac} = -\frac{1}{2}P^{(2)}_{ac}(\varepsilon_a + \varepsilon_c) = \overline{W}^{(2)}_{ca} \quad \forall\ a \in \text{vact},\ c \in \text{virt} \quad (15)$$

Occupied-Virtual Block (o–v)

$$\overline{W}^{(2)}_{ai} = -P^{(2)}_{ai}\varepsilon_i = -P^{(2)}_{ia}\varepsilon_i = \overline{W}^{(2)}_{ia} \quad \forall\ a \in \text{virt},\ i \in \text{occ} \quad (16)$$

Type III
Occupied–Occupied Block (o–o)

$$\overline{\overline{W}}^{(2)}_{ik} = -\frac{1}{2}\sum_{pq}^{MO}P^{(2)}_{pq}A_{pqki} = \overline{\overline{W}}^{(2)}_{ki} \quad \forall\ i,k \in \text{occ} \quad (17)$$

In these expressions, orbital energies (ε_p) are given as the eigenvalues of the converged Fock matrix, and $A_{aibj} \equiv 2(ai|bj) - (ab|ij) - (aj|ib)$ is an element of the orbital Hessian matrix, a Fock-like matrix provided by the coupled-perturbed Hartree-Fock (CPHF) formalism. To generate the o–v block of these RI-MP2 correction matrices, the Z-vector equation of Handy and Schaefer[64] is solved utilizing the following RI-MP2 specific Lagrangian (η_j and η_b will equal 0 for frozen orbitals and 1 for active orbitals within the occupied and virtual subspaces, respectively):

$$\begin{aligned} L_{bj} = {} & \eta_j\sum_{a}^{\text{vact}}\sum_{Q}^{\text{AUX}}(ba|Q)\Gamma^{Q}_{ja} - \eta_b\sum_{i}^{\text{act}}\sum_{Q}^{\text{AUX}}(ij|Q)\Gamma^{Q}_{ib} \\ & + \sum_{i}^{\text{act}}\sum_{K}^{\text{core}}P^{(2)}_{Ki}A_{bjKi} + \sum_{a}^{\text{vact}}\sum_{C}^{\text{vf}}P^{(2)}_{Ca}A_{bjCa} \\ & + \frac{1}{2}\sum_{ac}^{\text{vact}}P^{(2)}_{ac}A_{acbj} + \frac{1}{2}\sum_{ik}^{\text{act}}P^{(2)}_{ik}A_{ikbj} \end{aligned} \quad (18)$$

With these quantities defined, the analytical gradient at the MP2 level of theory has been presented in the general spin-orbital basis with explicit consideration of the active and frozen sectors of the correlation space and the RI approximation. In the next section, we present a theoretical formalism involving the RI-MP2 specific Lagrangian that allows for efficient construction of the quantities necessary for evaluation of this analytical gradient.

The Mixed Lagrangian Formalism

From the definition of the RI-MP2 Lagrangian given above in (18), we can now back-transform the first two terms into the mixed AO/MO basis as $L_{di}(1) = \sum_{\mu}^{\text{AO}} C_{\mu d}L_{\mu i}(1)$ and $L_{al}(2) = \sum_{\nu}^{\text{AO}} C_{\nu l}L_{a\nu}(2)$, where

$$L_{\mu i}(1) = \sum_{a}^{\text{vact}}\sum_{Q}^{\text{AUX}}(\mu a|Q)\Gamma^{Q}_{ia} = \frac{1}{2}\sum_{ab}^{\text{vact}}\sum_{j}^{\text{act}}(\mu a\|jb)t^{ab}_{ij} = L_{i\mu}(1) \quad (19)$$

and

$$L_{a\nu}(2) = -\sum_{i}^{\text{act}}\sum_{Q}^{\text{AUX}}(i\nu|Q)\Gamma^{Q}_{ia} = -\frac{1}{2}\sum_{ij}^{\text{act}}\sum_{b}^{\text{vact}}(i\nu\|jb)t^{ab}_{ij} = L_{\nu a}(2) \quad (20)$$

These terms are efficient to build from a computational standpoint and represent two core quantities that will be constructed in our RI-MP2 gradient algorithm. By utilizing this mixed Lagrangian formalism, we can now construct a number of quantities given in eqs. (6)–(17) via the following prescriptions:

$$P^{(2)}_{iK} = \sum_{\mu}^{\text{AO}} C_{\mu K}L_{\mu i}(1)/(\varepsilon_i - \varepsilon_K),$$

$$P^{(2)}_{aC} = -\sum_{\nu}^{\text{AO}} C_{\nu C}L_{a\nu}(2)/(\varepsilon_a - \varepsilon_C),$$

$$W^{(2)}_{ik} = -\sum_{\mu}^{\text{AO}} C_{\mu k}L_{\mu i}(1),$$

$$W^{(2)}_{ac} = \sum_{\nu}^{\text{AO}} C_{\nu c}L_{a\nu}(2),$$

and

$$W^{(2)}_{aj} = \sum_{\nu}^{\text{AO}} C_{\nu j}L_{a\nu}(2).$$

At this point, we can now identify the following set of core quantities, namely, Γ^{Q}_{ia}, Γ^{RS}, $P^{(2)}_{ki}$, $P^{(2)}_{ca}$, $L_{\mu i}(1)$, and $L_{a\nu}(2)$ that need to be constructed for evaluating the analytical RI-MP2 gradient. To proceed any further, specification of the HF wavefunction type is necessary. Therefore, in the next section, we specialize the results derived above in the general spin-orbital basis to handle both UHF and RHF wavefunctions.

Derivation of Wavefunction-Specific Quantities

To specifically derive the quantities necessary for both the UHF and RHF formalisms, we will integrate over the spin variables present in the core quantities described above. In all cases, beta subspace quantities are given by the interchange of α and β indices in the quantities derived specifically in the alpha subspace. Unless otherwise noted, either one of these quantities can be employed during post-RHF computation.

The RI-MP2 Specific Corrections to the 2-PDM

Contributions via the Three-Centered Forces: Γ^{Q}_{ia}. Integration over the spin variables in eq. (2) will yield the following Γ^{Q}_{ia} matrix in the alpha subspace:

$$\Gamma^{Q}_{i^\alpha a^\alpha} = \sum_{j^\alpha}^{\text{act}}\sum_{b^\alpha}^{\text{vact}} t^{a^\alpha b^\alpha}_{i^\alpha j^\alpha}C^{Q}_{j^\alpha b^\alpha} + \sum_{j^\beta}^{\text{act}}\sum_{b^\beta}^{\text{vact}} t^{a^\alpha b^\beta}_{i^\alpha j^\beta}C^{Q}_{j^\beta b^\beta} \quad (21)$$

For the RHF case,

$$\Gamma_{ia}^{\mathrm{Q,RHF}} = \Gamma_{i^\alpha a^\alpha}^{Q} + \Gamma_{i^\beta a^\beta}^{Q},$$

with the additional requirement that the alpha and beta spatial orbitals are equivalent, i.e.,

$$\Gamma_{ia}^{Q,\mathrm{RH}F} = 2\sum_{j^\alpha}^{\mathrm{act}}\sum_{b^\alpha}^{\mathrm{vact}} \tilde{t}_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} C_{j^\alpha b^\alpha}^{Q}, \tag{22}$$

where $\tilde{t}$, an RHF-specific t-amplitude, $\tilde{t}_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} \equiv (2(i^\alpha a^\alpha|j^\alpha b^\alpha) - (i^\alpha b^\alpha|j^\alpha a^\alpha))/\Delta_{i^\alpha j^\alpha}^{a^\alpha b^\alpha}$, has been introduced.

Contributions via the Two-Centered Forces: Γ^{RS}. Since the quantity given in eq. (3) for Γ^{RS} is a contraction of two spin-dependent quantities, there will also be two distinct Γ^{RS} matrices unique to RI-MP2 theory. In the alpha subspace, Γ^{RS} takes on the following form

$$\Gamma^{RS}(\alpha) = \sum_{i^\alpha}^{\mathrm{act}}\sum_{a^\alpha}^{\mathrm{vact}} C_{i^\alpha a^\alpha}^{R} \Gamma_{i^\alpha a^\alpha}^{S} \tag{23}$$

The RI-MP2 Corrections to the 1-PDM (P)

Active-Active Block: $P_{ik}^{(2)}$. From, eq. (7) we have $P_{ik}^{(2)} = -\frac{1}{2}\sum_{j}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}} t_{ij}^{ab} t_{kj}^{ab}$ which we will now transform to the equivalent expression $P_{kj}^{(2)} = -\frac{1}{2}\sum_{i}^{\mathrm{act}}\sum_{ab}^{\mathrm{vact}} t_{ij}^{ab} t_{ik}^{ab}$ for computational convenience (this expression will be utilized in the algorithm section). Integrating out spin, the RI-MP2 specific correction to the active–active block of the one-particle density matrix in the alpha subspace takes on the following form

$$P_{k^\alpha j^\alpha}^{(2)} = -\frac{1}{2}\sum_{i^\alpha}^{\mathrm{act}}\sum_{a^\alpha b^\alpha}^{\mathrm{vact}} t_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} t_{i^\alpha k^\alpha}^{a^\alpha b^\alpha} - \sum_{i^\beta}^{\mathrm{act}}\sum_{a^\beta b^\alpha}^{\mathrm{vact}} t_{i^\beta j^\alpha}^{a^\beta b^\alpha} t_{i^\beta k^\alpha}^{a^\beta b^\alpha} \tag{24}$$

and once again, interchange of spin indices will yield this quantity in the beta subspace. For the RHF case, $P_{kj}^{(2),\mathrm{RHF}} = P_{k^\alpha j^\alpha}^{(2)} + P_{k^\beta j^\beta}^{(2)}$, which yields

$$P_{kj}^{(2),\mathrm{RHF}} = -2\sum_{i^\alpha}^{\mathrm{act}}\sum_{a^\alpha b^\alpha}^{\mathrm{vact}} \tilde{t}_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} \frac{(i^\alpha a^\alpha|k^\alpha b^\alpha)}{\Delta_{i^\alpha k^\alpha}^{a^\alpha b^\alpha}} \tag{25}$$

upon relaxation of the UHF constraint on the delineation of the alpha and beta subspaces.

Vactive–Vactive Block: $P_{ca}^{(2)}$. The RI-MP2 correction to the vactive–vactive block of the one-particle density matrix in the alpha subspace takes on the following form after integrating out spin in eq. (8):

$$P_{c^\alpha a^\alpha}^{(2)} = \frac{1}{2}\sum_{i^\alpha j^\alpha}^{\mathrm{act}}\sum_{b^\alpha}^{\mathrm{vact}} t_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} t_{i^\alpha j^\alpha}^{c^\alpha b^\alpha} + \sum_{i^\alpha j^\beta}^{\mathrm{act}}\sum_{b^\beta}^{\mathrm{vact}} t_{i^\alpha j^\beta}^{a^\alpha b^\beta} t_{i^\alpha j^\beta}^{c^\alpha b^\beta} \tag{26}$$

For the RHF case, $P_{ca}^{\mathrm{RHF}} = P_{c^\alpha a^\alpha}^{(2)} + P_{c^\beta a^\beta}^{(2)}$, or

$$P_{ca}^{\mathrm{RHF}} = 2\sum_{i^\alpha j^\alpha}^{\mathrm{act}}\sum_{b^\alpha}^{\mathrm{vact}} \tilde{t}_{i^\alpha j^\alpha}^{a^\alpha b^\alpha} \frac{(i^\alpha c^\alpha|j^\alpha b^\alpha)}{\Delta_{i^\alpha j^\alpha}^{c^\alpha b^\alpha}} \tag{27}$$

where equivalence of the spatial extents of the alpha and beta subspaces was once again enforced due to the RHF formalism.

Parts (1) and (2) of the Lagrangian: $L_{\mu i}(1)$, $L_{a\nu}(2)$. In the alpha subspace, the first and second parts of the Lagrangian in the mixed AO/MO basis take on the following forms [c.f. eqs. (19), (20)]:

$$L_{\mu i^\alpha}(1) = \sum_{a^\alpha}^{\mathrm{vact}}\sum_{Q}^{\mathrm{aux}} (\mu a^\alpha|Q)\Gamma_{i^\alpha a^\alpha}^{Q} \tag{28}$$

and

$$L_{a^\alpha \nu}(2) = -\sum_{i^\alpha}^{\mathrm{act}}\sum_{Q}^{\mathrm{aux}} (i^\alpha \nu|Q)\Gamma_{i^\alpha a^\alpha}^{Q} \tag{29}$$

Now that we have formally specialized our spin-orbital expression of the first geometric derivative of the RI-MP2 correlation energy expression to accommodate both UHF and RHF wavefunctions, we begin our description of the computational effort required to evaluate the RI-MP2 analytical gradient. In the next section, we provide a detailed analysis of a computational algorithm that constructs and manipulates these aforementioned quantities and allows us to explore molecular potential energy surfaces at the RI-MP2 level of theory.

Algorithm

An overview of the RI-MP2 analytical gradient algorithm is given in Chart 1. In each of the following subsections, the individual functions in Chart 1 will be discussed in greater detail. In the following sections and the remainder of the paper, O, V, N, and X refer to the number of active occupied orbitals, active virtual orbitals, AO basis functions, and auxiliary basis functions, respectively.

Function 00: SCF Procedure

In *Function 00*, the standard SCF procedure is executed, generating the following set of matrices necessary during the post HF computation: the overlap (S), Fock (F), and density (P) matrices in the AO basis, and the coefficient (C) and orbital energy (ε) matrices in the MO basis. The computational cost associated with the SCF procedure and the associated memory and storage requirements are minimal in the context of our RI-MP2 gradient algorithm.

Functions 01-02: RI Overhead Routines

Function 01 is essentially one of two mandatory overhead routines for computations involving the RI approximation. In *Function 01*, the two-centered ERIs, $(P|Q)$ are assembled using the Coulomb metric in the auxiliary basis. A singular value decomposition (SVD) is then performed on $(P|Q)$ to obtain the desired inverse matrix necessary for constructing C, the RI-MP2 coefficient matrix described earlier. The computational cost of forming $(P|Q)^{-1}$ is cubic with respect to system size ($\sim X^3$) and is attributed entirely to the diagonalization of $(P|Q)$. The memory requirements associated with the diagonalization of $(P|Q)$, namely $3X^2$, also set the minimum overall memory requirements of any algorithm that employs the RI approximation. In our RI-MP2 gradient algorithm, we have decided on $3X^2$ as the overall memory requirement, thereby allowing gradient evaluation to be

Function 00	SCF Procedure; Formation of Overlap (S), MO Coefficient (C), Fock (F), Density (P), and Orbital Energy (ε) Matrices
Function 01	RI- Overhead (Part I): Formation of the $(P \mid Q)^{-1}$ Matrix
Function 02	RI- Overhead (Part II): Formation of the $(ia \mid P)$ Matrix
Function 03	Construction of the C_{ia}^{Q} Matrix
Function 04	Assembly of the Γ_{ia}^{Q}, $P_{ca}^{(2)}$, and $P_{kj}^{(2)}$ Matrices
Function 05	Construction of the Γ^{RS} Matrix
Function 06	Assembly of the L, P, and W Matrices; Solution of the Z -Vector Equations; Final Gradient Evaluation

Chart 1. A general flowchart of the RI-MP2 gradient algorithm.

performed with an equivalent memory cost to single-point energy evaluation. Therefore, with 2GB of available memory, our RI-MP2 gradient algorithm is able to treat system sizes on the order of 110 non-hydrogen atoms at the cc-pVDZ level. For future use, the $(P|Q)^{-1}$ matrix is written to disk with quadratic storage requirements (X^2).

In *Function 02*, the remainder of the overhead associated with the RI approximation is completed as the three-centered ERIs, $(ia|P)$, are constructed in the mixed MO/auxiliary basis. This set of three-centered ERIs is formed in a two-step transformation procedure from the pre-assembled set of three-centered ERIs in the mixed AO/auxiliary basis $(\mu\nu|P)$. In the first transformation step, $\mu \mapsto i$ via $(i\nu|P) = \sum_{\mu}^{\mathrm{AO}} C_{\mu i}\,(\mu\nu|P)$ with an associated quartic computational cost of $\sim O(nb2)X$ where $(nb2)$ is the number of significant AO basis set pairs (determined by the chosen integral threshold value). In the second transformation step, $\nu \mapsto a$ via $(ia|P) = \sum_{\nu}^{\mathrm{AO}} C_{\nu a}\,(i\nu|P)$ with an associated quartic computational cost of $\sim NOVX$. The set of $(ia|P)$ is also written to disk with cubic storage requirements (OVX) for use in later functions.

Function 03: Construction of the C_{ia}^{Q} Matrix

In *Function 03*, the RI-MP2 coefficient matrix is constructed via matrix multiplication over auxiliary basis set functions as given by eq. (4) with an associated quartic computational cost of $\sim OVX^2$. Since the set of three-centered integrals, $(ia|P)$, was written to disk in *Function 02* with the overall order $[a,P,i]$, i.e., a is the fastest index and i is the slowest index, the C matrix can be most efficiently assembled in a loop over i. For a given i, a subset of the $(ia|P)$ matrix $\forall\ a \in$ vact, $P \in$ aux is loaded from disk, and contracted with the full set of two-centered integrals, $(P|Q)^{-1}$, which is held in memory throughout the routine. As the C matrix is constructed for a given i, it is also written to disk with the same order and overall storage requirements as the three-centered integrals (OVX).

Function 04: Assembly of the Γ_{ia}^{Q}, $P_{ca}^{(2)}$, and $P_{kj}^{(2)}$ Matrices

Function 04 delicately manages the construction of all quantities requiring quintic computational effort. The manner in which we proceed uses a fairly involved batching scheme that carefully balances CPU and wall timings under the working memory constraint of $3X^2$, cubic disk storage, quartic I/O requirements, and quintic computational cost. Without this complex batching scheme, our algorithm was susceptible to dominance by I/O time due to an unmanageable number of hard drive seeks. Given that the $(ia|P)$ and C_{ia}^{Q} arrays are written to disk with the aforementioned order, we suggest the algorithm depicted in Chart 2.

The batch sizes involved in the loops over active occupied and active virtual batches are determined by the overall memory constraint of $3X^2$ and are dynamically calculated for each molecular system considered (to maximize the size of each batch). In this function, the steps requiring quintic computational effort are completed as incremental matrix multiplications to maximize flop efficiency. These steps include the construction of the full set of four-centered ERIs, $(ia|jb)$, and the core matrices necessary to evaluate the analytical gradient, namely Γ_{ia}^{Q}, $P_{ca}^{(2)}$, and $P_{kj}^{(2)}$. The computational costs associated with the generation of these quantities are $\sim O^2V^2X$, $\sim O^2V^2X$, $\sim O^2V^3$, and $\sim O^3V^2$, respectively. Unlike the algorithm used to compute the single-point energy at the RI-MP2 level of theory, we were unable to take advantage of the fact that $(ia|jb)$ is inherently lower triangular in the gradient evaluation. To exploit this fact computationally, which would allow us to only compute approximately half of the total O^2V^2 elements, our algorithm would necessitate several additional quartic I/O steps (or quartic disk requirements)—sacrifices that would prove to be too costly in the large-molecule limit.

In *Function 04*, three quartic I/O steps are needed to construct the core quantities described above. To form the set of integrals, whether or not the lower triangular structure of $(ia|jb)$ is exploited, requires the acquisition of C_{jb}^{Q} from disk with an associated I/O cost of O^2VX. In our current implementation, we

```
Loop over active occupied orbitals, $i$
  Load $(ia \mid P)\ \forall\ a, P, given\ i$ from disk                                $[a, P]$

    Loop over batches of active occupied orbitals, $ob$

        Loop over $j \in ob$
          Load $C_{jb}^{P}\ \forall\ b, P, given\ j$ from disk                          $[b, P]$
          Make $(ia \mid jb) = \sum_P (ia \mid P) C_{jb}^{P}\ \forall\ a, b, given\ (ij)$    $[a, b]$
          Make $t_{ij}^{ab} = (ia \parallel jb) \big/ \Delta_{ij}^{ab}\ \forall\ a, b, given\ (ij)$   $[a, b]$
          Accumulate $t_{ij}^{ab}\ \forall\ a, b, j \in ob, given\ i$                     $[a, b, j \in ob]$
          Increment $E_{RI-MP2} += \frac{1}{4} t_{ij}^{ab} (ia \parallel jb)$
          Increment $P_{ca} += \sum_b t_{ij}^{ab} t_{ij}^{cb}\ \forall\ a, c, given\ (ij)$   $[a, c]$
          Increment $\Gamma_{ia}^{P} += \sum_b t_{ij}^{ab} C_{jb}^{P}\ \forall\ a, P, given\ (ij)$   $[a, P]$
        End Loop over $j \in ob$

        Loop over batches of active virtual orbitals, $vb$
          Extract $t_{ij}^{ab}\ \forall\ a \in vb, b, j \in ob, given\ i$               $[a \in vb, b, j \in ob]$
          Write $t_{ij}^{ab}\ \forall\ a \in vb, b, j \in ob, given\ i$ to disk         $[a \in vb, b, j]$ (HD)
        End Loop over batches of active virtual orbitals, $vb$

    End Loop over batches of active occupied orbitals, $ob$

  Write $\Gamma_{ia}^{P}\ \forall\ a, P, given\ i$ to disk                              $[a, P, i]$ (HD)

    Loop over batches of active virtual orbitals, $vb$
      Load $t_{ij}^{ab}\ \forall\ a \in vb, b, j, given\ i$                             $[a \in vb, b, j]$

        Loop over $a \in vb$
          Extract $t_{ij}^{ab}\ \forall\ b, j, given\ (ia)$                             $[b, j]$
          Increment $P_{kj} += \sum_b t_{ik}^{ab} t_{ij}^{ab}\ \forall\ j, k, given\ (ia)$   $[j, k]$
        End Loop over $a \in vb$

    End Loop over batches of active virtual orbitals, $vb$

End Loop over active occupied orbitals, $i$
```

Chart 2. A more detailed look at the algorithm in *Function 04*. Array and disk (HD) ordering is given with index speeds decreasing from left to right within [square brackets].

have not encountered systems where this quartic I/O step has manifested itself as a marked discrepancy between CPU and wall timings. The remaining quartic I/O steps are associated with the reading and writing of the various t-amplitude matrices to and from disk. By contrast with the original RI-MP2 gradient algorithm presented by Weigend and Häser,[56] where amplitudes are read from disk in a process accompanied by a large number of hard drive seeks, we suggest the batching routine depicted in Chart 2. In this routine, the number of data elements transferred during communication with the hard drive is unchanged, namely O^2V^2, but the number of hard drive seeks is dramatically reduced from $(O^2 + O^2V)$ to $(ON_{\nu b} + ON_{ob}N_{\nu b})$ where N_{ob} and $N_{\nu b}$ are the number of active occupied and active virtual batches, respectively. Therefore, the fundamental scaling associated with the number of seeks is left unchanged, but the prefactor is significantly reduced. To illustrate the importance of reducing the number of hard drive seeks in this algorithm, consider the plots depicted in Figure 1. From these plots, it is clear that batching is critical to avoid being dominated by the I/O time associated with random access seeks. Alternative approaches to this potential I/O

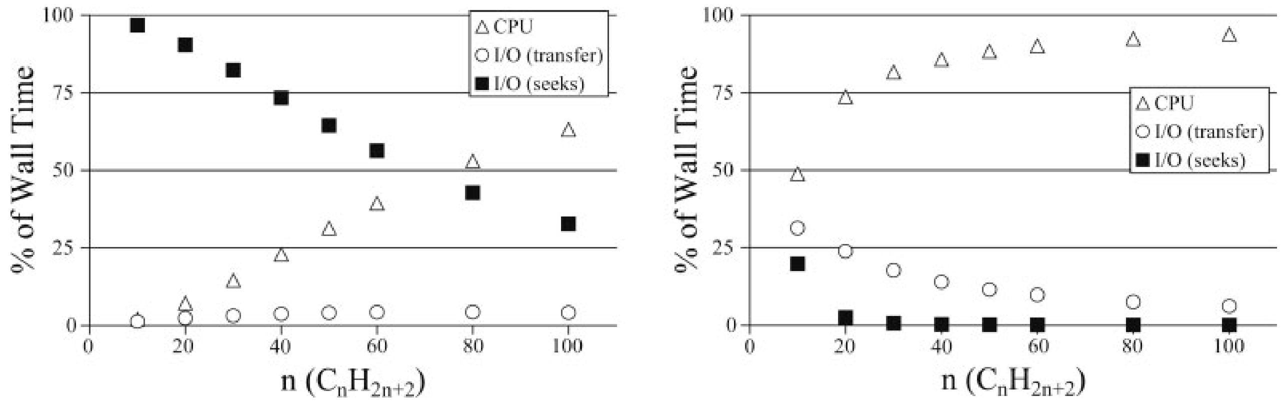

Figure 1. Theoretical plot of % of wall time vs. n (number of carbon atoms in an alkane chain) for CPU, I/O transfer, and I/O seeks for the algorithm presented in the original RI-MP2 gradient algorithm[56] (left panel) and in this work (right panel). To generate these plots, the following parameters were used: processor speed of 4.0 Gflops, 100 MB/s data transfer rate, and 10 ms per hard drive seek. Batch sizes were maximized for each system considered based on the overall memory constraint of $3X^2$. Cache efficiency was not taken into consideration during the construction of this figure.

bottleneck include a direct approach,[62] in which the *t*-amplitudes are formed as needed; however, this scheme results in additional quintic computational steps that may fare as prohibitive in the large molecule limit.

Function 05: Construction of the Γ^{RS} Matrix

Function 05 serves the purpose of assembling Γ^{RS} via the prescription given by eq. (3) with an associated computational cost of $\sim OVX^2$. The structure of this routine is quite simple as the Γ^R_{ia} and C^S_{ia} matrices are read from disk within a governing loop over active occupied orbitals, *i*. The desired quantity, Γ^{RS}, is constructed via incremental matrix multiplication and written to disk with quadratic storage requirements.

Function 06: Assembly of the L, P, and W Matrices; Solution of the Z-Vector Equations; Final Gradient Evaluation

Function 06 is actually a series of small routines that contribute to the evaluation of the RI-MP2 analytical gradient. To begin, Γ^Q_{ia} is back-transformed from the MO basis to the mixed MO/AO basis ($\Gamma^Q_{ia} \mapsto \Gamma^Q_{i\nu}$) with an associated computational cost of $\sim OVNX$ and is saved to disk with cubic storage requirements (*ONX*). Next, the mixed Lagrangian formalism is invoked as $L_{\mu i}(1)$ and $L_{\nu a}(2)$ are assembled using both Γ matrices (which are still stored on disk) as follows:

$$L_{\mu i}(1) = \sum_{\nu}^{\text{AO}} \sum_{Q}^{\text{aux}} (\mu\nu|Q)\Gamma^Q_{i\nu} \tag{30}$$

$$L_{\nu a}(2) = -\sum_{i}^{\text{act}} \sum_{Q}^{\text{aux}} (i\nu|Q)\Gamma^Q_{ia} \tag{31}$$

The computational costs for the construction of these mixed Lagrangian matrices are both quartic, namely, $\sim O(nb2)X$ and $\sim OVNX$, respectively. Note that the construction of these quantities requires that the three-centered ERIs must again be assembled in the AO basis and transformed into the mixed MO/AO basis with an additional quartic CPU cost of $\sim O(nb2)X$.

Next, the RI-MP2 specific correction to the 2-PDM is once again back-transformed from the mixed MO/AO basis to the AO basis ($\Gamma^Q_{i\nu} \mapsto \Gamma^Q_{\mu\nu}$ with a CPU cost of $\sim ON^2X$) and is contracted with the set of three-centered DERIs, $\{(\mu\nu|Q)^x\}$, providing the first complete term in the analytical gradient for a cubic computational cost of $\sim(nb2)X$. For a quadratic number of flops ($\sim X^2$), the two-centered forces are computed by contracting Γ^{RS} with the set of two-centered DERIs, $\{(R|S)^x\}$, providing another complete term.

The *L, P,* and *W* matrices are assembled within several different subroutines that deal specifically with the frozen core and frozen virtual approximations as described by the definitions given above. To complete the Lagrangian assembly, the RI-MP2 correction to the 1-PDM (*P*) must be contracted with the series of integrals generated by CPHF theory (the *A* matrices) with an associated quartic cost of $(nb2)^2$ (this is done in the AO basis for increased efficiency, which will cost several additional but negligible cubic computational transformation steps)—a process that is analogous to the construction of the Fock matrix.

To generate the o–v block of the RI-MP2 correction to the 1-PDM, the Z-vector equations are solved via conjugate gradient methods with a cost comparable to the prerequisite SCF procedure. The RI-MP2 analytical gradient is then completed by the remaining contractions: the 1-PDM ($P^{\text{SCF}}_{\mu\nu}$), the energy-weighted 1-PDM ($W^{\text{SCF}}_{\mu\nu}$), and the 2-PDM ($\Gamma^{\text{SCF}}_{\mu\nu\lambda\sigma}$) from SCF theory are added to their respective RI-MP2 corrected quantities (i.e., $P^{(2)}_{\mu\nu}$, $W^{(2)}_{\mu\nu}$, and $\Gamma^{(2)}_{\mu\nu\lambda\sigma}$) and are then contracted with the appropriate derivative quantities, namely, $H^{(x)}_{\mu\nu}$, $S^{(x)}_{\mu\nu}$, and $\{(\mu\nu|\lambda\sigma)^x\}$, respectively. These final steps have associated quadratic ($\sim(nb2)$ for contractions involving the Hamiltonian and overlap derivatives) and quartic computational cost ($\sim(nb2)^2$ for contractions involving the four-centered integral derivatives).

Summary of Computational Requirements

As mentioned above, the RI-MP2 analytical gradient algorithm was written under the following constraints: quadratic memory not to exceed $3X^2$, cubic disk (storage) requirements, quartic I/O transfer, and quintic computational cost. Within the six functions described above, there are four steps with quintic computational cost. These dominant computational steps are attributed to the formation of the full set of four-centered ERIs, $(ia|jb)$, and the core matrices necessary to evaluate the analytical gradient, namely Γ^{Q}_{ia}, $P^{(2)}_{ca}$, and $P^{(2)}_{kj}$. The computational cost associated with the generation of these quantities is $\sim O^2V^2X$, $\sim O^2V^2X$, O^2V^3, and $\sim O^3V^2$, respectively. For comparison, only construction of the four-centered ERIs is necessary for single-point energy evaluation at the RI-MP2 level of theory. During the analytical gradient evaluation, four cubic quantities are written to disk, namely, $(ia|Q)$, C^{Q}_{ia}, Γ^{Q}_{ia}, and $\Gamma^{Q}_{i\mu}$, in such a manner that the overall storage requirements are $OVX + ONX$, a quantity only marginally larger than the required allocation of $2OVX$ in RI-MP2 energy computation. Both energy and gradient evaluations require quartic data transfer as the RI-MP2 coefficient matrix is read into core memory within the inner loop over occupied orbitals in *Function 04*. However, the gradient algorithm necessitates the additional transfer of the *t*-amplitudes (O^2V^2 data elements) within the same loop. One additional key feature of this algorithm is that its in-core memory requirements are identical to those utilized in single-point calculations. This feature allows us to evaluate analytical RI-MP2 gradients on any system for which the RI-MP2 energy is feasible.

Algorithm Performance

All calculations discussed in this work were performed on a single 2 GHz AMD Opteron processor with 2 GB available memory and 100 GB available disk space. The RI-MP2 analytical gradient code was implemented into the Q-Chem 3.0[65] software package, which was used for all calculations in this work. All DFT computations performed throughout this work utilized the popular hybrid B3LYP[66,67] functional. Therefore, all references to the DFT methodology hereafter specifically refer to the B3LYP variant. During all post-HF computations, the frozen-core approximation was introduced to reduce computational timings and resources. Unless otherwise specified, self-consistency was determined by DIIS errors below 10^{-8} a.u. and the numerical threshold utilized in all integral evaluations was set at 10^{-12} a.u. Geometry optimizations were converged when two of the following three criteria were achieved: a maximum energy change of 1.0×10^{-6} hartrees, a maximum force/gradient component of 3.0×10^{-4} hartrees/bohr, and a maximum displacement of 1.2×10^{-3} bohr. The AO basis sets employed in this work include the correlation consistent basis sets [cc-pVXZ where X = D, T, and Q] of Dunning[68] and the polarized double- and triple-ζ quality sets of Ahlrichs [VDZ(d)[69] and TZVPP[70]] and Pople [6-31G(d), 6-31G(d,p), 6-311G(2df), 6-311G(2df,2pd)].[71–75] In this work, we have chosen to utilize the complementary auxiliary basis sets provided by Weigend et al.[44,76] (available via ftp://ftp.chemie.uni-karlsruhe.de/pub/cbasen/) since the size of these expansion sets are smaller than those of Bernholdt and Harrison.[77]

In the Basis Set Selection Section, we address the use of Pople-style basis sets in RI-MP2 energy evaluations by computing atomization energies of the G3/99 set[78–80] relative to conventional MP2 theory. In the Computational Timings Section, we first consider the effects of varying the parameters utilized in determining SCF convergence and screening during integral evaluation on the timing and numerical stability of our RI-MP2 analytical gradient algorithm. We further explore the performance of our RI-MP2 analytical gradient algorithm by providing a detailed comparative analysis of the computational timings associated with single-point energy evaluation and analytical force determination of porphyrin using the HF, DFT, RI-MP2, and MP2 methodologies and several different basis sets of double- and triple-ζ quality. For completeness, in the Numerical Assessment Section, the performance of our RI-MP2 analytical gradient will be numerically assessed against experimental data and the predictions of conventional MP2 theory as an entire set of 136 test molecules are optimized using various basis sets of double- and triple-ζ quality, including several representative Pople-style sets.

Basis Set Selection

When performing computations of double- and triple-ζ quality, one is faced with the choice of several different basis sets—the most common choices, of course, include cc-pVDZ and cc-pVTZ, the correlation-consistent basis sets of Dunning, and the analogously polarized Pople-style sets, 6-31G(d,p) and 6-311G(2df,2pd). Often when using Pople-style basis sets, the polarization functions on the hydrogen atoms are omitted during molecular geometry optimizations to achieve noteworthy speedups while introducing only minor differences in the final optimized geometries (e.g. via the 6-31G(d) and 6-311G(2df) basis sets). At the present time, however, the Pople-style AO basis sets do not have specially optimized complementary auxiliary basis sets, and for that reason, they are not often utilized in RI-MP2 calculations. However, there are auxiliary basis sets currently available that are designed to accompany the double-ζ polarization (DZP) quality VDZ(d) and cc-pVDZ AO basis sets-sets that are very similar in structure to the commonly used 6-31G(d) and 6-31G(d,p) basis sets. Therefore, we expect that the currently available auxiliary basis sets, when paired with the Pople-style AO basis sets, would still provide us with relatively accurate results during RI-MP2 computations.

To test this assumption, we first computed atomization energy errors (relative to MP2 theory) in the G3/99 set using the RI-MP2 methodology and the following complementary AO/auxiliary basis set pairs: VDZ(d)/aux-VDZ(d) and cc-pVDZ/aux-cc-pVDZ (double-ζ level); TZVPP/aux-TZVPP and cc-pVTZ/aux-cc-pVTZ (triple-ζ level) to numerically quantify the errors associated with the use of complementary AO/auxiliary basis set pairs in RI-MP2 single-point energy evaluations. With these errors in hand, we then paired the Pople-style basis sets of DZP quality, 6-31G(d) and 6-31G(d,p), with both the aux-VDZ(d) and aux-cc-pVDZ auxiliary basis sets and computed the relative atomization energies of the G3/99 set. The analogous procedure was also carried out at the triple-ζ level, where the 6-311G(2df) and 6-311G(2df,2pd) basis sets were employed in combination with the aux-TZVPP and aux-cc-pVTZ auxiliary basis sets.

At the DZP level, we found root-mean-squared (RMS) relative errors of 0.138 and 0.087 kcal/mol for the complementary

Table 1. Relative Atomization Energy Errors (in kcal/mol) for the G3/99 Set.

Level	AO basis	AUX basis	RMS	MAX
Double-ζ	VDZ(d)	aux-VDZ(d)	0.138	0.466
	cc-pVDZ	aux-cc-pVDZ	0.087	0.309
	6-31G(d)	aux-VDZ(d)	0.179	0.838
		aux-cc-pVDZ	0.184	0.496
	6-31G(d,p)	aux-VDZ(d)	0.222	0.838
		aux-cc-pVDZ	0.226	0.627
Triple-ζ	TZVPP	aux-TZVPP	0.071	0.189
	cc-pVTZ	aux-cc-pVTZ	0.059	0.155
	6-311G(2df)	aux-TZVPP	0.075	0.503
		aux-cc-pVTZ	0.061	0.155
	6-311G(2df,2pd)	aux-TZVPP	0.081	0.529
		aux-cc-pVTZ	0.064	0.167

These relative errors were computed by directly comparing the atomization energies predicted by the MP2 and RI-MP2 methodologies using the AO basis sets listed above (and auxiliary basis sets when applicable). The above statistics do not include computations on the Li_2 and Na_2 systems, where the aux-TZVPP auxiliary basis set was not adequately equipped with higher angular momentum functions to approximate the basis function pairs in the triple-ζ Pople-style AO basis sets.

VDZ(d)/aux-VDZ(d) and cc-pVDZ/aux-cc-pVDZ AO/auxiliary basis set pairs, respectively (Table1). When the Pople-style 6-31G(d) [6-31G(d,p)] AO basis sets were paired with the aux-VDZ(d) and aux-cc-pVDZ auxiliary basis sets, the RMS errors were computed as 0.179 [0.222] kcal/mol and 0.184 [0.226] kcal/mol, respectively. Although the magnitude of these errors are larger than those for the complementary pairs, one should realize that these relative atomization energy errors (~0.1 to 0.2 kcal/mol) are well within the requirements for chemical accuracy (generally considered to be approximately 1 kcal/mol for atomization errors) and completely negligible when compared with the atomization energy itself—a quantity that is typically hundreds or thousands of kilocalories per mole. Based on these findings, we now feel justified in pairing these AO basis sets with either auxiliary basis set during RI-MP2 energy computations. Since the RMS relative errors are slightly smaller for computations employing the aux-VDZ(d) basis set, and the fact that the aux-VDZ(d) set (1458 functions) is marginally smaller than the aux-cc-pVDZ set (1540 functions) for porphyrin, we have chosen to pair the Pople-style AO basis sets with the aux-VDZ(d) auxiliary basis set in the following computational timings analysis.

At the triple-ζ level, complementary AO/auxiliary basis set pairs generated reference RMS relative errors of 0.071 kcal/mol (TZVPP/aux-TZVPP) and 0.059 kcal/mol (cc-pVTZ/aux-cc-pVTZ); as expected, the magnitude of these relative errors decreased as the quality of the AO and auxiliary basis sets was increased. When the Pople-style 6-311G(2df) [6-311G(2df,2pd)] AO basis sets were paired with the aux-TZVPP and aux-cc-pVTZ auxiliary basis sets, the RMS errors were computed as 0.075 [0.081] kcal/mol and 0.061 [0.064] kcal/mol, respectively. Even though the RMS relative errors are smaller in magnitude when the more complete auxiliary basis set is employed, these errors are so small when compared with atomization energies that we feel comfortable using either of these auxiliary basis sets in conjunction with these Pople-style AO basis sets. During the subsequent timings analysis in Computational Timings Section, we chose to pair the smaller aux-TZVPP set with the Pople-style AO basis sets (2244 in aux-TZVPP vs. 2364 in aux-cc-pVTZ).

Computational Timings

To begin exploring the performance of the algorithm presented in this work, we first turn our attention to the effects of varying the parameters utilized in determining SCF convergence (**CONV**) and screening during integral evaluation (**THRESH**) on the computational timings and numerical stability associated with analytical force determinations (a single step during a geometry optimization) on porphyrin at the RI-MP2/cc-pVDZ level of theory (the Cartesian coordinates for this system are provided in the supplementary material). For reference, $10^{-\mathbf{CONV}}$ is the value of the DIIS error vector utilized in determining SCF convergence and $10^{-\mathbf{THRESH}}$ is the threshold value, which sets the drop tolerance for general integral evaluation and AO shell pair overlap (the quantity referred to as *nb2* throughout this work), respectively. Based on previous experience, we have chosen the parameters 8 (**CONV**) and 12 (**THRESH**) for use in the Chemical Application section that follows, but suspect that these values may be slightly more conservative than necessary for our purposes.[81]

In Table2, we present the relative force and energy errors introduced by various **CONV/THRESH** combinations accompanied by the relative computational time necessary for each analytical force determination. For clarity in Table 2, we have not presented the analogous MP2/cc-pVDZ computations, as those findings were almost identical to the RI-MP2/cc-pVDZ results. In this analysis we have deemed pairs of **CONV/THRESH** values that introduce relative force errors that are approximately one order of magnitude less than the aforementioned convergence criterion (3.0×10^{-4} hartrees/bohr) to constitute acceptable alternatives to the benchmark 8/12 standards. Considering the maximum (MAX) relative force errors in Table 2, all analytical force determinations using **CONV** values of 6, 7, and 8 met the relative force error criterion described above, i.e., the following **CONV/THRESH** combinations were found to be acceptable alternatives to the 8/12 benchmark: 6/8, 6/9, 6/10; 7/9, 7/10, 7/11; 8/10, 8/11. Taking energy into consideration, the total energy errors introduced within the series of **CONV/THRESH** combinations discussed above only ranges from 2 to 14 microhartrees. Therefore, using the 6/8 combination, one can not only reproduce the results of an analytical force computation at the 8/12 level, but one can do so in approximately half the time without changing the energy to within a hundredth of a kcal/mol!

To justify these relative timings, one needs to first consider the savings at the SCF level, where smaller **CONV** values directly reduce the number of iterations required to generate a self-consistent field while lower **THRESH** values minimize the computational effort needed to assemble the Fock matrix by truncating the number of retained integrals and AO shell pairs. During the post-HF portion of the analytical force determination, the effect of varying the **CONV** parameter on the CPU timings is negligible. Instead, the lower **THRESH** values predominantly reduce the CPU time by speeding up the iterative solution of the CPHF equations and the contraction of the separable 2-PDM cor-

Table 2. The Effects of SCF Convergence (**CONV**) and Integral Threshold (**THRESH**) Parameters on the RI-MP2 Analytical Gradient Algorithm Presented in this Work Using a Single 2 GHz AMD Opteron Processor.

		Force errors		Energy errors			Relative CPU time		
CONV	**THRESH**	RMS	MAX	SCF	CORR	TOTAL	SCF	CORR	TOTAL
5	7	0.000073	0.000187	0.000308	0.000089	0.000397	21.2	57.3	44.6
	8	0.000061	0.000174	−0.000011	0.000118	0.000107	27.6	64.8	51.7
	9	0.000061	0.000173	−0.000003	0.000118	0.000115	34.3	73.5	59.7
6	8	0.000002	0.000007	−0.000019	0.000005	−0.000014	31.4	64.9	53.0
	9	0.000002	0.000006	–	0.000002	0.000003	39.2	71.9	60.4
	10	0.000002	0.000007	–	0.000002	0.000002	48.3	79.2	68.3
7	9	–	0.000001	0.000002	–	0.000002	44.6	72.3	62.5
	10	–	0.000001	–	–	–	55.5	79.3	70.9
	11	–	0.000001	–	–	–	67.7	89.1	81.6
8	10	–	–	–	–	–	67.0	79.3	75.0
	11	–	–	–	–	–	82.1	91.9	88.4
	12	–	–	–	–	–	100.0	100.0	100.0

On the left, RMS and MAX force component errors are presented in hartrees/bohr. Mean-field (SCF), correlation (CORR), and the combined mean-field + correlation (TOTAL) energy errors are given in hartrees. Relative CPU timings (in %) for the iterative SCF procedure (SCF), all post-HF computation (CORR), and the total analytical force determination (TOTAL) are shown on the right. All quantities presented are based on comparisons to analytical force determination using 8 and 12 as the values for **CONV** and **THRESH**, respectively. Hyphenated entries in the table above were computed to be zero at the level of accuracy presented.

rection with the four-centered AO-DERIs—both of these processes are analogous to the aforementioned Fock matrix construction in that their associated computational costs rely heavily upon sparsity amongst AO basis functions. For completeness, we chose to vary the convergence criteria utilized during the iterative CPHF equations (usually set at 6) in tandem with the 8/12 combination discussed above. However, even the mildest relaxation of the convergence criteria introduced an unacceptable MAX relative force error of approximately 4.6×10^{-4} hartrees/bohr and is therefore not recommended during routine geometry optimizations.

With these findings in mind, we now explore the relationship between computational timings and basis set size for the HF, DFT, MP2, and RI-MP2 methodologies. In Figure 2, we present the computational timings associated with single-point energy evaluations and analytical force determinations on porphyrin using these methods and the aforementioned AO basis sets (and auxiliary basis sets when applicable). For clarity, we chose to omit DFT from the plots as the CPU time for DFT is almost always slightly larger (but on the same order of magnitude) than HF. For both single-point energy evaluations and analytical force determinations at the RI-MP2 and MP2 levels, the CPU timings presented in this figure include the time necessary for the prerequisite iterative SCF procedure—all timings presented in Figure 2 reflect the total time necessary to perform each computation from scratch.

Among the series of double- and triple-ζ quality basis sets, the RI approximation clearly provides us with the ability to obtain second-order perturbative energy corrections to the HF energy at very little additional cost over the underlying SCF procedure. For example, with respect to the HF computational timings using the same basis set, the additional time necessary for RI-MP2 [MP2] energy evaluations at the 6-31G(d), 6-31G(d,p), and cc-pVDZ levels was found to be 34.1 [168.33]%, 33.8 [180.3]%, and 11.9 [86.6]%, respectively. Although post-HF computations appear to be more efficient when the cc-pVDZ basis set is employed, this result is merely a by-product of the fact that the SCF procedure using the cc-pVDZ basis set is tremendously inefficient. In fact, the CPU time necessary to generate a self-consistent field using the 6-31G(d,p) basis set was more than two times faster than the same computation using the cc-pVDZ basis set, despite the fact that the 6-31G(d,p) basis set is slightly larger than cc-pVDZ! Although each of these basis sets includes the same number of *s* and *p* functions on each heavy atom in porphyrin (3*s* and 2*p* for C, N, and O), only the Pople-style basis sets utilize *sp* shells (where *s* and *p* functions share common exponents but have different contraction coefficients). This basis set structure can be exploited in AO integral evaluation algorithms that loop over AO shells, such as those utilized in the Q-Chem software package.[65] Therefore, the computational savings observed when SCF calculations are performed using the Pople-style basis set are a direct result of rapid AO integral evaluation during construction of the Fock matrix. Note, of course, that the computational bottleneck during RI-MP2 energy evaluation is the formation of the $(ia|jb)$ integrals via matrix multiplication—a fifth-order step which does not explicitly depend on AO integral evaluation.

When performing analytical force determinations, one quickly realizes that the RI approximation offers smaller speedups than had previously been accomplished during single-point energy evaluations. Nevertheless, the use of the RI approximation in MP2 theory still reduces the overall associated computational cost and thereby extends the current size limitations to practical MP2 computations. At the double- and triple-ζ levels, for instance, the post-HF time required to evaluate the RI-MP2 analytical gradient was 2.52 [2.22], 2.42 [2.39], and 2.00 [2.08] times faster than the analogous MP2 computations using the 6-31G(d) [6-311G(2df)], 6-31G(d,p) [6-311G(2df,2pd)], and cc-pVDZ [cc-pVTZ] basis sets. As seen

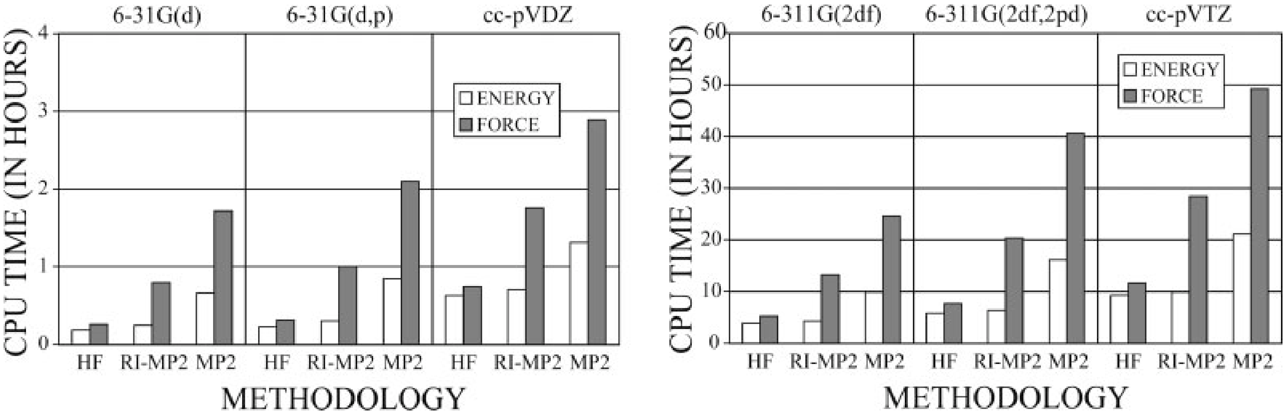

Figure 2. Plots of CPU time (in hours) vs. methodology (HF, RI-MP2, and MP2) utilized during single-point energy evaluations and analytical force determinations of porphyrin on a single 2 GHz AMD Opteron processor. In each of the panels, the effect of employing different basis sets of double-ζ (left panel) and triple-ζ (right panel) quality on the computational timings was explored. All timings presented reflect the total time necessary to perform each computation from scratch. The number of AO basis functions employed in the 6-31G(d) [6-311G(2df)], 6-31G(d,p) [6-311G(2df,2pd)], and cc-pVDZ [cc-pVTZ] single-point energy evaluations and analytical force determinations was 388 [762], 430 [916], and 406 [916], respectively. The number of functions in the auxiliary basis sets was 1458 [aux-VDZ(d)], 1540 [aux-cc-pVDZ], 2244 [aux-TZVPP], and 2364 [aux-cc-pVTZ].

above when investigating the CPU timings associated with energy evaluations, even greater savings are attained when the Pople-style basis sets are employed; RI-MP2 analytical force determinations using the 6-31G(d,p) [6-311G(2df)] and 6-31G(d) [6-311G(2df, 2pd)] basis sets were found to be 1.76 [1.40] and 2.22 [2.15] times faster than the analogous cc-pVDZ [cc-pVTZ] computations, respectively. This trend can be explained by the increased computational efficiency during AO integral evaluation and the contraction of the separable 2-PDM correction with the four-centered AO-DERIs resulting from the use of *sp* shells in the Pople-style sets.

These findings represent a marked improvement over current MP2 gradient implementations in that a single geometry optimization step at the double-ζ [triple-ζ] level for molecular systems the size of a porphyrin can now be accomplished in just under 25 min [7 h] on a single desktop processor. However, these noteworthy speedups over conventional MP2 theory are dependent on system size. When treating systems of this size ($\sim$20 to 25 heavy atoms), the major SCF-related steps (SCF, Z-vector equation, and AO-DERI contractions) which scale as $\sim(nb2)^2$ still dominate the fifth order matrix multiplication steps necessary for RI-MP2 analytical gradient evaluation. Indeed this is why the speedups are not larger. These fifth order steps begin to dominate for systems that include more than 40 heavy atoms, and in this regime of molecular system size, RI-MP2 analytical force determinations show even larger speed-ups of approximately 2.6–3.0 over conventional MP2 theory at the double-ζ level (based on test calculations using alanine octapeptide and alanine hexadecapeptide).

Numerical Assessment: The Equilibrium Experimental Test Set

In this section, we present a brief comparative analysis of the performance of the RI-MP2 and MP2 methodologies in the geometry optimization of 136 test molecules, which includes 1st and 2nd row atoms (i.e., from H-Ar) and ranges from H_2 to benzene in size, for which experimentally determined equilibrium bond lengths (r_e values) are available for each of the included 166 symmetry-unique bonds (Table3). Of these 136 test molecules, 92 of them are closed-shell systems (including 3 anions, 77 neutrals, and 12 cations) while the remaining 44 require the unrestricted formalism (including 2 anions, 30 neutrals, and 12 cations). We shall refer to this test set as the Equilibrium Experimental Test Set (EXTS). By performing geometry optimizations using the aforementioned methodologies and a series of different basis sets, we are now in a position to numerically assess the performance of the RI-MP2 analytical gradient against experimental findings and the theoretical predictions of conventional MP2 theory. In addition, we hope to provide further insight into the use of Pople-style AO basis sets with the currently available double- and triple-ζ auxiliary basis sets during RI-MP2 geometry optimizations.

The mean (MAD) and maximum (MAX) absolute deviations from experimentally-determined structural data during the optimization of EXTS are provided in Table 4. At the double- and triple-ζ levels, we found that the RI-MP2 and MP2 methodologies were in complete agreement in all cases to within 0.1 pm, the bond distance that is generally considered to define the realm of chemical accuracy in molecular equilibrium structure determinations. From this analysis, we concluded that RI-MP2 computations using the double-ζ Pople-style AO basis sets coupled with the aux-VDZ(d) and aux-cc-pVDZ auxiliary basis sets slightly outperformed the same computations utilizing the complementary cc-pVDZ/aux-cc-pVDZ AO/auxiliary basis set pair by 0.006 Å (6-31G(d)) and 0.008 Å (6-31G(d,p)). This trend was duplicated at the triple-ζ level, where RI-MP2 computations that paired either the aux-TZVPP or aux-cc-pVTZ auxiliary basis sets with the Pople-style AO basis sets yielded fractionally better results than the analogous cc-pVTZ/aux-

Table 3. The 136 Molecules Included in the Equilibrium Experimental Test Set (EXTS) and the Corresponding 166 Symmetry-Unique Experimentally-Determined Bond Lengths (in Å).

MOLECULAR SYSTEMS INCLUDED IN EXTS							
Molecule	r_e	Molecule	r_e	Molecule	r_e	Molecule	r_e
C_2^- ($^2\Sigma_g^+$)	1.268[a]	CO_2	1.160[b]	LiF ($^1\Sigma^+$)	1.564[a,b]	CH_2O_2 (*trans*)	1.091 (CH)[c]
NH^- ($^2\Pi_i$)	1.047[a]	CS ($^1\Sigma^+$)	1.535[a,b]	LiH ($^1\Sigma^+$)	1.596[a]		1.201 (C=O)[c]
AlH^+ ($^2\Sigma^+$)	1.602[a]	CS_2	1.553[c]	MgO ($^1\Sigma^+$)	1.749[a]		1.340 (CO)[c]
Cl_2^+ ($^2\Pi_{3/2g}$)	1.892[a]	F_2 ($^1\Sigma_g{}^+$)	1.412[a]	MgS ($^1\Sigma^+$)	2.143[a]		0.969 (OH)[c]
CO^+ ($^2\Sigma^+$)	1.115[a]	F_2O	1.405[b]	N_2 ($^1\Sigma_g{}^+$)	1.098[a,b]	CHF_3	1.091 (CH)[c]
HCl^+ ($^2\Pi_i$)	1.315[a]	F_2S	1.587[c]	NaCl ($^1\Sigma^+$)	2.361[a,b]		1.328 (CF)[c]
He_2^+ ($^2\Sigma_u^+$)	1.081[a]	F_2Si	1.590[b]	NaF ($^1\Sigma^+$)	1.926[a,b]	ClCN	1.160 (CN)[b,c]
HF^+ ($^2\Pi_i$)	1.001[a]	H_2 ($^1\Sigma_g{}^+$)	0.741[a,b]	NaH ($^1\Sigma^+$)	1.887[a]		1.629 (CCl)[b,c]
N_2^+ ($^2\Sigma_g^+$)	1.116[a]	SiN ($^2\Sigma^+$)	1.572[a]	NF ($^3\Sigma^-$)	1.317[b,a]	F_3HSi	1.562 (SiF)[b]
NH^+ ($^2\Pi_r$)	1.070[a]	H_2N^-	1.028[c]	NH ($^3\Sigma^-$)	1.036[a,b]		1.447 (SiH)[b]
$NH_3{}^+$	1.014[c]	NO^- ($^3\Sigma^-$)	1.258[a]	NH_3	1.012[b]	H_2CCCH_2	1.308 (CC)[c]
$O_2{}^+$ ($^2\Pi_g$)	1.116[a]	PO^- ($^3\Sigma^-$)	1.540[a]	O_2 ($^3\Sigma_g{}^-$)	1.208[a,b]		1.076 (CH)[c]
PF^+ ($^2\Pi_r$)	1.500[a]	BeH^+ ($^1\Sigma^+$)	1.312[a]	O_3	1.272[b,c]	H_2CS	1.611 (CS)[c]
H_2O^+ (^2B_l)	0.999[c]	CH^+ ($^1\Sigma^+$)	1.131[a]	P_2 ($^1\Sigma_g{}^+$)	1.893[a,b]		1.086 (CH)[c]
AlS ($^2\Sigma^+$)	2.029[a]	CN^+ ($^1\Sigma^+$)	1.173[a]	PF ($^3\Sigma^-$)	1.590[a]	HCCCN	1.062 (CH)[c]
BeCl ($^2\Sigma^+$)	1.797[a]	$H_3{}^+$	0.877[c]	PH ($^3\Sigma^-$)	1.422[a]		1.206 (C≡C)[c]
BeF ($^2\Sigma^+$)	1.361[a]	H_3O+	0.976[c]	PH_3	1.413[c]		1.376 (CC)[c]
BeH ($^2\Sigma^+$)	1.343[a]	MgH^+ ($^1\Sigma^+$)	1.652[a]	PN ($^1\Sigma^+$)	1.491[a,b]		1.161 (C≡N)[c]
BO ($^2\Sigma^+$)	1.205[a,b]	NO^+ ($^1\Sigma^+$)	1.063[a]	S_2 ($^3\Sigma_g{}^-$)	1.889[a,b]	HCCH	1.061 (CH)[b]
BS ($^2\Sigma^+$)	1.609[a]	NS^+ ($^1\Sigma^+$)	1.440[a]	SCS	1.553[b]		1.203 (CC)[b]
CF ($^2\Pi_r$)	1.272[a,b]	OH^+ ($^3\Sigma^-$)	1.029[a]	SiO ($^1\Sigma^+$)	1.510[a,b]	HCN	1.065 (CH)[b,c]
CH ($^2\Pi_r$)	1.120[a]	SiH^+ ($^1\Sigma^+$)	1.504[a]	SiS ($^1\Sigma^+$)	1.929[a,b]		1.153 (CN)[b,c]
CH3	1.076[c]	$NH_4{}^+$	1.021[c]	SO ($^3\Sigma^-$)	1.481[a,b]	HCP	1.066 (CH)[c]
ClO ($^2\Pi_i$)	1.570[a]	AlCl ($^1\Sigma^+$)	2.130[a,b]	SO_2	1.431[b,c]		1.540 (CP)[c]
CN ($^2\Sigma^+$)	1.172[a]	AlF ($^1\Sigma^+$)	1.654[a,b]	HOO (2A”)	0.971 (OH)[c]	HNC	0.994 (NH)[c]
CP ($^2\Sigma^+$)	1.562[a]	AlH ($^1\Sigma^+$)	1.648[a]		1.331 (OO)[c]		1.169 (NC)[c]
H_2N ($^2B_1$)	1.025[c]	AlN ($^3\Pi_i$)	1.786[a]	HCO (2A’(Π))	1.119 (CH)[c]	HOCl	0.964 (OH)[c]
HO	0.970[a]	BCl ($^1\Sigma^+$)	1.797[a]		1.175 (CO)[c]		1.689 (OCl)[c]
MgCl ($^2\Sigma^+$)	2.199[a]	BeO ($^1\Sigma^+$)	1.331[a]	HCO^+	1.105 (CO)[c]	N_2O	1.127 (NN)[c]
MgF ($^2\Sigma^+$)	1.750[a]	BeS ($^1\Sigma^+$)	1.742[a]		1.097 (CH)[c]		1.185 (NO)[c]
MgH ($^2\Sigma^+$)	1.730[a]	BF ($^1\Sigma^+$)	1.263[a,b]	C_6H_6	1.086 (CH)[c]	OCS	1.147 (CO)[b]
NO ($^2\Pi_r$)	1.151[a,b]	BF_3	1.307[c]		1.390 (CC)[c]		1.561 (CS)[b]
NO_2 ($2A_l$)	1.195[c]	BH ($^1\Sigma^+$)	1.232[a]	CCl_2O	1.177 (CO)[c]	S_2O	1.884 (SS)[c]
NS ($^2\Pi_r$)	1.494[a,b]	BN ($^3\Pi$)	1.281[a,b]		1.737 (CCl)[c]		1.456 (SO)[c]
OP ($^2\Pi_r$)	1.474[b]	CF_4	1.315[c]	CH_2Cl_2	1.080 (CH)[c]	B_2H_6	1.184 (BH ×4)[c]
SH ($^2\Pi_i$)	1.345[b]	CH_4	1.087[c]		1.766 (CCl)[c]		1.314 (BH ×2)[c]
SiCl ($^2\Pi_r$)	2.058[a]	H_2O	0.958[b,c]	CH_2F_2	1.084 (CH)[c]	cyclopropane	1.501 (CC)[c]
SiF ($^2\Pi_r$)	1.601[a]	H_2S	1.336[b,c]		1.351 (CF)[c]		1.083 (CH)[c]
SiH ($^2\Pi_r$)	1.520[a,b]	H_2Si	1.514[c]	CH_3Cl	1.778 (CCl)[b]	CCl ($^2\Pi_{1/2}$, $^2\Pi_{3/2}$)	1.645[a,b]
Cl_2 ($^1\Sigma_g^+$)	1.988[a,b]	HCl ($^1\Sigma^+$)	1.275[a,b]		1.086 (CH)[b]	SF ($^2\Pi_{3/2}$, $^2\Pi_{1/2}$)	1.601[a]
ClF	1.628[a,b]	HF ($^1\Sigma^+$)	0.917[a,b]	CH_3F	1.383 (CF)[c]		
CO ($^1\Sigma^+$)	1.128[a,b]	LiCl ($^1\Sigma^+$)	2.021[a,b]		1.086 (CH)[c]		

[a] Ref. 82(a).
[b] Ref. 82(b).
[c] Ref. 82(c).

cc-pVTZ computations by approximately 0.002 Å. These findings, coupled with the atomization energy results in the Basis Set Selection Section and the computational timings results in the Computational Timings Section, strongly suggest that reliable RI-MP2 equilibrium structures can be generated at a substantially reduced computational cost when the Pople-style AO basis sets are used in tandem with the currently available double- and triple-ζ auxiliary basis sets during RI-MP2 geometry optimizations.

Chemical Application

When dealing with systems of biochemical interest, size constraints almost always limit theoretical research to molecular mechanics and semi-empirical computations. In turn, these methodologies often depend on parameters derived from structural and energetics data obtained using high-level quantum mechanical computations on model systems. In this section, we utilize our

Table 4. Mean (MAD) and Maximum (MAX) Absolute Deviations (in Å) from Experimentally Determined Structural Data Found During Optimization of EXTS.

		B3LYP		MP2		RI-MP2		
Level	AO basis	MAD	MAX	MAD	MAX	MAD	MAX	AUX basis
Double-ζ	cc-pVDZ	0.021	0.068	0.024	0.067	0.024	0.068	aux-cc-pVDZ
	6-31G(d)	0.015	0.054	0.018	0.080	0.018	0.080	aux-VDZ(d)
		–	–	–	–	0.018	0.080	aux-cc-pVDZ
	6-31G(d,p)	0.014	0.054	0.016	0.080	0.016	0.080	aux-VDZ(d)
		–	–	–	–	0.016	0.080	aux-cc-pVDZ
Triple-ζ	cc-pVTZ	0.010	0.072	0.012	0.080	0.012	0.080	aux-cc-pVTZ
	6-311G(2df)	0.009	0.073	0.010	0.081	0.010	0.081	aux-TZVPP
		–	–	–	–	0.010	0.081	aux-cc-pVTZ
	6-311G(2df,2pd)	0.008	0.073	0.010	0.081	0.010	0.081	aux-TZVPP
		–	–	–	–	0.010	0.081	aux-cc-pVTZ

These relative errors were computed by directly comparing the experimentally determined bond lengths (r_e) with the predictions of the B3LYP, RI-MP2, and MP2 methodologies using the AO basis sets listed above (and auxiliary basis sets when applicable). Note that the mean absolute deviations presented above were computed utilizing the following prescription: $\Sigma(r_{\text{theory}} - r_e)/N$ with $N = 166$ (to account for each of the symmetry-unique bonds).

RI-MP2 analytical gradient code to characterize the gap between extended and globular conformations of alanine tetrapeptide. In doing so, we will assess the performance of RI-MP2 theory against the HF and DFT methods in both the geometry optimizations and single-point energy evaluations of two alanine tetrapeptide conformations (one extended and one globular). Beginning with initial geometries optimized at the HF/6-31G(d,p) level of theory by Friesner and coworkers,[83] one extended conformation (hereby referred to as **E**, corresponding to conformation 1 in ref. 53) and one globular conformation (**G**, corresponding to conformation 3 in ref. 53) of alanine tetrapeptide, i.e., NME-ALA1-ALA2-ALA3-ACE, were optimized using the HF, B3LYP, and RI-MP2 methodologies and the cc-pVDZ and cc-pVTZ basis sets (the Cartesian coordinates of these optimized geometries are provided in the supplementary material). Following geometry optimization, single-point energy evaluations were performed on all structures using the aforementioned methodologies and the cc-pVXZ (X = D, T, Q) basis set series to characterize the energetics associated with the aforementioned extended-globular conformational gap (the SCF, B3LYP, and RI-MP2 energies of these optimized geometries are provided in the supplementary material). Unless otherwise noted, the following results and discussion will be based on the structures optimized using the more complete cc-pVTZ basis set.

Using the HF and B3LYP methodologies, single-point energy calculations were performed on all optimized structures to generate the relative energy gaps depicted in Table5. In general, HF single-point energy evaluations that included higher angular momentum functions, i.e., at the cc-pVTZ and cc-pVQZ levels, incorrectly predicted that **E** was more stable than **G** for the conformations optimized at all levels of theory. In fact, we observed an overall relative destabilization of **G** (or stabilization of **E**) by 2.722 kcal/mol [HF], 2.681 kcal/mol [B3LYP], and 3.939 kcal/mol [RI-MP2] as the basis set quality was increased from cc-pVDZ to cc-pVQZ during HF single-points. Similar results were observed when the B3LYP variant of DFT was employed; as the basis set was systematically increased using the cc-pVXZ series, we observed an overall relative destabilization of **G** by 3.080, 3.154, and 4.958 kcal/mol among the structures optimized using the HF, B3LYP, and RI-MP2 methodologies, respectively. In both of these cases, most of the observed changes occurred at the intermediate cc-pVTZ level, which strongly suggests that the relatively large degree of intramolecular BSSE present when incomplete basis sets are employed, i.e., at the cc-pVDZ level, can be held accountable for this trend. The changes in these relative conformational energy gaps as the basis set quality was increased from cc-pVTZ to cc-pVQZ ranged from 0.5 to 1.0 kcal/mol, indicating that the inclusion of higher angular momentum functions in HF and B3LYP single-point energy evaluations was significant for these systems, as computations at the cc-

Table 5. Relative Energy Gaps (in kcal/mol) for a Series of Two Different Alanine Tetrapeptide Conformations.

Optimization level	HF/cc-pVXZ			B3LYP/cc-pVXZ		
	D	T	Q	D	T	Q
HF/cc-pVDZ	0.714	−1.632	−2.254	3.010	0.237	−0.475
B3LYP/cc-pVDZ	−0.163	−2.931	−3.627	3.493	0.073	−0.776
RI-MP2/cc-pVDZ	−1.257	−4.460	−5.263	2.925	−1.061	−2.055
HF/cc-pVTZ	0.666	−1.497	−2.056	2.520	0.079	−0.560
B3LYP/cc-pVTZ	0.465	−1.700	−2.216	3.061	0.485	−0.093
RI-MP2/cc-pVTZ	−1.364	−4.513	−5.303	3.638	−0.303	−1.320

These relative energy gaps were computed as the energy difference between conformations **E** and **G**, *i.e.*, ΔE = Energy(**E**)–Energy(**G**). The relative energy gaps were generated by single-point energy calculations using the HF and B3LYP methodologies and the cc-pVXZ series (where X = D, T, or Q) on structures optimized using the HF, B3LYP, and RI-MP2 methodologies and the cc-pVDZ (upper panel) and cc-pVTZ (lower panel) basis sets. The number of atomic orbital basis functions employed in the cc-pVDZ, cc-pVTZ, and cc-pVQZ single-point energy evaluations was 390, 908, and 1760, respectively.

Table 6. Relative Energy Gaps (in kcal/mol) for a Series of Two Different Alanine Tetrapeptide Conformations.

Optimization level	RI-MP2/cc-pVXZ				
	D	T	D→T	Q	T→Q
HF/cc-pVDZ	6.382	4.644	4.900	3.835	3.698
B3LYP/cc-pVDZ	7.759	6.111	6.584	5.198	5.039
RI-MP2/cc-pVDZ	8.054	6.036	6.535	4.983	4.801
HF/cc-pVTZ	5.378	3.727	3.942	3.003	2.884
B3LYP/cc-pVTZ	6.411	5.186	5.582	4.522	4.414
RI-MP2/cc-pVTZ	8.001	6.167	6.720	5.156	4.994

These relative energy gaps were computed as the energy difference between conformations **E** and **G**, *i.e.*, ΔE = Energy(**E**)–Energy(**G**). Single-point energy computations were performed at the RI-MP2/cc-pVXZ (where X = D, T, or Q) level of theory (using the respective complementary auxiliary basis sets) on structures optimized using the HF, B3LYP, and RI-MP2 methodologies and the cc-pVDZ (upper panel) and cc-pVTZ (lower panel) basis sets to yield the relative energy gaps depicted above. Data in the D→T and T→Q columns represent energy gaps at the extrapolated RI-MP2/cc-pV(DT)Z and RI-MP2/cc-pV(TQ)Z levels, respectively. The number of atomic orbital (auxiliary) basis functions employed in the cc-pVDZ, cc-pVTZ, and cc-pVQZ single-point energy evaluations was 390 (1428), 908 (2260), and 1760 (3580), respectively.

pVQZ level were necessary to obtain reasonably converged approximations to the complete basis set limit.

In Table 6, the relative energy gaps generated by single-point energy evaluations using the RI-MP2 methodology are given. At the RI-MP2/cc-pVDZ level, the relative conformational energy gap was found to be 5.378, 6.411, and 8.001 kcal/mol for the HF, B3LYP, and RI-MP2 optimized structures, respectively. In going from cc-pVDZ to cc-pVTZ, we observed a systematic decrease in the energy gaps, namely, 1.651 kcal/mol [HF], 1.225 kcal/mol [B3LYP], and 1.834 kcal/mol [RI-MP2]—a trend again directly attributable to intramolecular BSSE. As the quality of the basis set employed was increased to cc-pVQZ, the energy gaps were once again decreased by a fairly consistent amount but to a lesser degree, namely, 0.724 kcal/mol [HF], 0.664 kcal/mol [B3LYP], and 1.011 kcal/mol [RI-MP2], indicating steady convergence of the correlation energy. At the RI-MP2/cc-pVQZ level, **G** was found to be more stable than **E** by 3.003, 4.522, and 5.156 kcal/mol for the HF, B3LYP, and RI-MP2 optimized structures, respectively. At this level, the energy gaps derived from the B3LYP and RI-MP2 optimized structures were reasonably similar, differing by 0.634 kcal/mol, while the energy gap characterized by the HF structures was smaller by 1.519 kcal/mol [B3LYP] and 2.153 kcal/mol [RI-MP2]. These findings strongly suggest that certain structural features are absent when geometry optimizations that only account for polarization and electrostatics are utilized.

Since all single-point energy calculations utilized the correlation consistent basis sets of Dunning, we now have the option of utilizing the following prescription[1]

$$E_{XY} = E_{\mathrm{SCF},Y} + \frac{X^3 E_{\mathrm{CORR},X} - Y^3 E_{\mathrm{CORR},Y}}{X^3 - Y^3} \qquad Y > X \qquad (32)$$

where, $E_{\mathrm{SCF},Y}$ and $E_{\mathrm{CORR},Y}$ represent the SCF and correlation energy at the cc-pVYZ level of theory, respectively, to yield the relative conformational energies at the cc-pV(DT)Z and cc-pV(TQ)Z extrapolated levels (Table 6). At the cc-pV(DT)Z level, the extended-globular energy gaps were slightly increased from the cc-pVTZ level, with increases of 0.215 kcal/mol [HF], 0.396 kcal/mol [B3LYP], and 0.553 kcal/mol [RI-MP2]. At the more complete cc-pV(TQ)Z level, the extended-globular energy gaps were further decreased from the cc-pVQZ results by a slight amount, namely, 0.119 kcal/mol [HF], 0.108 kcal/mol [B3LYP], and 0.162 kcal/mol [RI-MP2]. These findings clearly demonstrate that T→Q extrapolation is necessary to ensure convergence to the sub-kcal/mol level of accuracy, a result also seen in our previous work[52] investigating the relative energetics among alanine tetrapeptide conformations. The extended-globular energy gaps at the extrapolated RI-MP2/cc-pV(TQ)Z level should be considered our "best" numbers and were computed as 2.884, 4.414, and 4.994 kcal/mol for the HF, B3LYP, and RI-MP2 optimized structures, respectively. As seen above at the cc-pVQZ level, the relative energy gaps predicted using the B3LYP and RI-MP2 optimized structures were still reasonably similar, differing by 0.580 kcal/mol, whereas the HF gap was notably distinct from the B3LYP [1.530 kcal/mol] and RI-MP2 [2.110 kcal/mol] cases.

Based on the extrapolated RI-MP2/cc-pV(TQ)Z data set, we have concluded that the energy gaps generated by the B3LYP and RI-MP2 optimized structures were approximately 1.5–2.1 kcal/mol larger than the gap generated using the HF optimized structures. These findings suggest that the conformers optimized using the B3LYP and RI-MP2 methodologies were more similar in structure to each other than their respective HF optimized analogs. To numerically quantify these structural differences, we computed the root-mean-squared structural deviations (RMSD) amongst all optimized conformers following Rhee.[84] For conformers **E** and **G**, the RMSD between the B3LYP and RI-MP2 optimized structures were computed as 0.317 and 0.284 Å, respectively. However, when comparing the HF and B3LYP [RI-MP2] optimized structures, instances in which we expect larger structural deviations, the RMSD values were only found to be 0.397 [0.152] Å and 0.112 [0.306] Å, for conformations **E** and **G**, respectively. To put these values into perspective, one should consider the RMSD between **E** and **G**, two conformations that are markedly different, computed as 3.717 Å [HF], 3.782 Å [B3LYP], and 3.929 Å [RI-MP2]. Based on these findings, we concluded that the RMSD results were inconclusive and warranted a more detailed examination of the optimized structures for further analysis.

Upon closer inspection of **E**, the linear alanine tetrapeptide conformation, we found that all of the optimized structures contained three unique intramolecular hydrogen bonds between carbonyl oxygens and amide hydrogens on the same residues. These hydrogen bond types were denoted as (1): ALA1 ⋯ ALA1,(2) ALA2 ⋯ ALA2, and (3) ALA3 ⋯ ALA3 (Fig. 3). In the HF optimized structure, the N-H⋯O hydrogen bond lengths [associated NHO bond angles] were found as 2.216 Å [104.71°], 2.216 Å [104.69°], and 2.224 Å [104.76°], for each of the aforementioned hydrogen bond types, respectively, placing them in the range of weak (<5 kcal/mol) interactions.[86] In comparing the quantities generated by the B3LYP [RI-MP2] optimized structures with those characterizing the HF structures, we noted a slight decrease in the

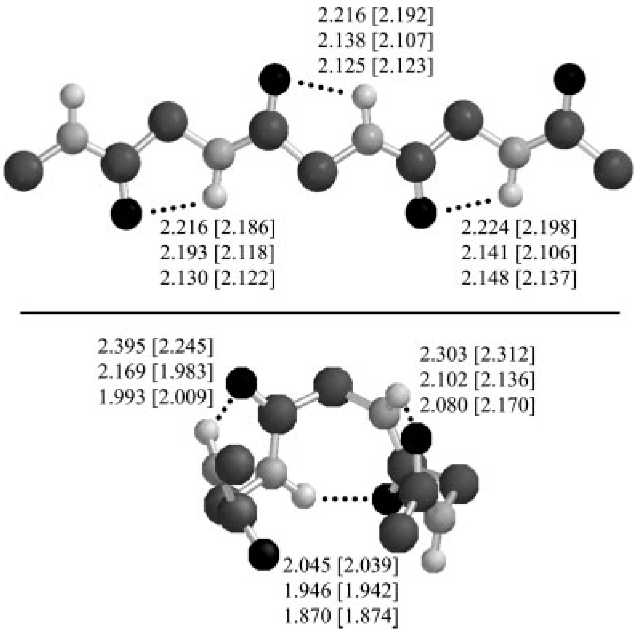

Figure 3. Intramolecular hydrogen bonding present in conformations **E** (top panel) and **G** (bottom panel). In each of these panels, the following grey-scale scheme was utilized for identification of the atoms in each alanine tetrapeptide backbone: hydrogen (white), nitrogen (light grey), carbon (dark grey), and oxygen (black). The hydrogen bond lengths are given in Å and represent the HF (top), B3LYP (middle), and RI-MP2 (bottom) optimized structures at the cc-pVTZ [cc-pVDZ] level. These figures were created using the Spartan '04 package.[85]

hydrogen bond lengths, 0.023 [0.086] Å, 0.078 [0.091] Å, and 0.083 [0.076] Å, accompanied by a slight increase in the associated bond angles, 1.49 [3.01]°, 3.13 [3.31]°, and 3.55 [2.85]°. Both of these small shifts suggest hydrogen bonds of slightly increased strength in the B3LYP and RI-MP2 optimized structures. The changes in bond lengths [angles] amongst the B3LYP and RI-MP2 optimized structures were smaller, namely, 0.063 Å [1.52°], 0.013 Å [0.04°], and 0.007 Å [0.70°]; in each of these cases (except the last), the RI-MP2 hydrogen bond lengths were slightly shorter and associated bond angles larger than in the B3LYP analog. Not only do these findings support the claim that the B3LYP and RI-MP2 optimized structures were more similar to one another than the HF optimized conformation, but they also suggest that extended B3LYP and RI-MP2 conformations are slightly more stable than the extended HF conformation. However, these findings alone are not sufficient enough to account for the energy gap discrepancy predicted by the T→Q extrapolation scheme above. To explain this energetic difference, we then performed a similar analysis on the globular alanine tetrapeptide conformation.

In **G**, we also found three distinct intramolecular hydrogen bonds in all of the optimized structures. However, these intramolecular hydrogen bonds were found between carbonyl oxygens and amide hydrogens on different residues, denoted as: (1) NME···ALA2, (2) ALA1···ALA3, and (3) ALA2···ACE (Fig. 3). In the HF optimized structure, the hydrogen bond lengths [angles] were found as 2.395 Å [128.09°], 2.045 Å [148.83°], and 2.303 Å [134.23°], for each of the aforementioned hydrogen bond types. When compared with the B3LYP [RI-MP2] optimized structures, we observed a marked decrease in the hydrogen bond lengths, 0.226 [0.402] Å, 0.099 [0.175] Å, and 0.201 [0.223] Å, accompanied by an increase in the associated bond angles, 6.14 [13.25]°, 2.06 [4.82]°, and 3.81 [3.16]°. With hydrogen bond lengths decreasing 0.175 [0.267] Å on average, placing them in the range of weak-to-moderate (5–15 kcal/mol) interactions,[86] it seems reasonable to expect that the B3LYP and RI-MP2 globular conformations will be stabilized to a much greater extent than the HF globular conformation. In addition, we observed that the B3LYP and RI-MP2 globular conformations were found to be less similar than their linear analogs. More specifically, we found that the hydrogen bond of type (1) was significantly shorter in the RI-MP2 optimized structures by 0.176 Å [accompanied by a bond angle increase of 7.11°]. Combining these findings with the analysis of the extended conformations, we now feel that the energetic discrepancies found at the RI-MP2/cc-pV(TQ)Z level can be justified. From the analysis of **E**, we qualitatively concluded that the RI-MP2 and B3LYP conformers were slightly more stable than the HF structure (RI-MP2 $\approx$ B3LYP $<$ HF). When the globular conformations were investigated, we observed a marked discrepancy among the degree of hydrogen bonding within each of the optimized structures. As a result, the relative energetic orderings maintained that the RI-MP2 optimized structure was stabilized to a much larger extent than the B3LYP and HF analogs (RI-MP2 $<$ B3LYP $\ll$ HF).

With these findings in mind, we are now in a position to comment further on the use of the HF, B3LYP, and RI-MP2 methodologies in generating the structural and energetic data necessary for parameterization of empirically-based computational methods. HF energetics account for polarization and electrostatics alone, with no explicit electron correlation treatment; in these cases, the extended conformation would clearly be favored over the globular conformation in alanine tetrapeptide. If the predictions of B3LYP energy evaluations were considered, short-range correlation would also be included. However, the resultant force field would fail to include explicit consideration of long-range van der Waals interactions and would also favor the extended conformation in most cases. These results are consistent with our previous work[52] where we found that MP2, when compared with B3LYP, provide both convergent and reliable relative conformational energies. This observation can be explained by the significant role that dispersion interactions play in the relative conformational energies of polypeptides.[87] In fact, correlation has been found to be critical in determining the relative stability among alanine and glycine dipeptides, where inclusion of higher-order correlation effects (i.e., from CCSD(T)) provided no significant difference over the predictions of MP2 theory.[88] Based on these findings and our current work, we feel justified in suggesting the use of the relative conformational energies generated by single-point computations at the RI-MP2 level of theory as the basis for parameterization of energetic quantities; in particular, we strongly recommend the use of the T→Q extrapolation scheme described above.

Structural data, on the other hand, was expected to be far less sensitive to the level of theory employed for geometry optimization. As a result of this analysis, however, we have found surprisingly significant changes in the optimized geometries as a function of level of theory, which in turn, have directly translated to chemically significant shifts in the computed energy difference between the extended and globular conformations of alanine tetrapeptide. In fact, single-point energy calculations at the RI-MP2/cc-pV(TQ)Z level attribute differences of 2.1 kcal/mol between the HF and RI-MP2 optimized structures and 0.6 kcal/mol between the B3LYP and RI-MP2 optimized structures. These differences clearly demonstrate the importance of long-range dispersion interactions in conformational analysis and are substantial enough to warrant recommending the routine use of RI-MP2 for geometry optimizations of small biomolecules when feasible.

Conclusions

In this work we reported a new algorithm that utilizes a semi-direct batching approach to minimize the potential I/O bottlenecks associated with disk-based storage and access of the RI-MP2 *t*-amplitudes during RI-MP2 analytical gradient evaluation. Using this algorithm, we have obtained computational speed-ups of approximately 2–3 over conventional MP2 analytical gradient algorithms. The dominant computational steps for systems (less than 40 heavy atoms) are in fact SCF-related steps (SCF iterations, Z-vector iterations, AO-DERI contractions), rather than true fifth order RI-MP2 steps. Therefore, in future work we hope to obtain further speedups in those SCF steps by utilizing density fitting[89] and/or a dual-basis approach.[90,91]

Based on 136 optimized geometries, as well as G3/99 database energetics, we found that RI-MP2 computations employing Pople-style AO basis sets and the existing auxiliary basis sets of Weigend et al.[44,76] can reproduce benchmark MP2 energies and structures, while offering considerable efficiency advantages in integral evaluation. In the chemical application of this algorithm, we computed the extended-globular conformational energy gap in alanine tetrapeptide at the extrapolated RI-MP2/cc-pV(TQ)Z level as 2.884, 4.414, and 4.994 kcal/mol for structures optimized using the HF, B3LYP, and RI-MP2 methodologies, respectively. The surprisingly large energetic consequences associated with using differentially optimized structures, coupled with the fact that RI-MP2 theory was the only method to consistently predict that the globular structure was lower in energy, clearly demonstrate the importance of long-range dispersion interactions in polypeptide conformational analysis. These findings strongly support the use of RI-MP2 structures and energetics for consistent and reliable parameterization of force fields and other semi-empirical methods designed to accurately treat large biochemical systems of interest.

Acknowledgments

We gratefully acknowledge useful discussions with Prof. Jay Ponder and Prof. Pengyu Ren, who suggested us to re-examine the conformational energies of the alanine tetrapeptides and forwarded the optimized structures to us. M. H.-G. is a part-owner of Q-Chem Inc.

References

1. Helgaker, T.; Jørgensen, P.; Olsen, J. Molecular Electronic Structure Theory; Wiley: Chichester, 2000.
2. Jensen, F. Introduction to Computational Chemistry; Wiley: Chichester, 1999.
3. Levine, I. Quantum Chemistry; Prentice Hall: Upper Saddle River, 2000.
4. Dunning, T. H. J Phys Chem A 2000, 104, 9062.
5. Helgaker, T.; Ruden, T. A.; Jørgensen, P.; Olsen, J.; Klopper, W. J Phys Org Chem 2004, 17, 913.
6. Coriani, S.; Marchesan, D.; Gauss, J.; Hättig, C.; Helgaker, T.; Jørgensen, P. J Chem Phys 2005, 123, 184107.
7. Szabo, A.; Ostlund, N. S. Modern Quantum Chemistry: An Introduction to Advanced Electronic Structure Theory; Dover: New York, 1989.
8. Pulay, P. Adv Chem Phys 1987, 69, 241.
9. Møller, C.; Plesset, M. S. Phys Rev 1934, 46, 618.
10. Head-Gordon, M.; Pople, J. A.; Frisch, M. J. Chem Phys Lett 1988, 153, 503.
11. Purvis, G. D.; Bartlett, R. J. J Chem Phys 1982, 76, 1910.
12. Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. Chem Phys Lett 1989, 157, 479.
13. Bartlett, R. J.; Watts, J. D.; Kucharcki, S. A.; Noga, J. Chem Phys Lett 1990, 165, 513.
14. Parr, R. G.; Yang, W. Density Functional Theory of Atoms and Molecules; Oxford University Press: New York, 1989.
15. Kohn, W.; Becke, A. D.; Parr, R. G. J Phys Chem 1996, 100, 12974.
16. Koch, W.; Holthausen, M. C. A Chemist's Guide to Density Functional Theory; Wiley-VCH: New York, 2000.
17. Head-Gordon, M. J Phys Chem 1996, 100, 13213.
18. Greengard, L. Science 1994, 265, 909.
19. White, C. A.; Head-Gordon, M. J Chem Phys 1996, 104, 2620.
20. White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. Chem Phys Lett 1996, 253, 268.
21. Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Science 1996, 271, 51.
22. Challacombe, M.; Schwegler, E. J Chem Phys 1997, 106, 5526.
23. Schwegler, E.; Challacombe, M.; Head-Gordon, M. J Chem Phys 1997, 106, 9708.
24. Ochsenfeld, C.; White, C. A.; Head-Gordon, M. J Chem Phys 1998, 109, 1663.
25. Scuseria, G. E. J Phys Chem A 1999, 103, 4782.
26. Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. J Chem Phys 1997, 106, 9639.
27. Johnson, B. G.; Gonzales, C. A.; Gill, P. M. W.; Pople, J. A. Chem Phys Lett 1994, 221, 100.
28. Perdew, J. P.; Zunger, A. Phys Rev B 1981, 23, 5048.
29. Kristyan, S.; Pulay, P. Chem Phys Lett 1994, 229, 175.
30. Pople, J. A.; Binkley, J. S.; Seeger, R. Int J Quantum Chem Symp 1976, 10, 1.
31. Hehre, W. J.; Radom, L.; Schleyer, P. V. R.; Pople, J. A. *Ab Initio* Molecular Orbital Theory; Wiley: New York, 1986.
32. Helgaker, T.; Gauss, J.; Jørgensen, P.; Olsen, J. J Chem Phys 1997, 106, 6430.
33. Byrd, E. F. C.; Sherrill, C. D.; Head-Gordon, M. J Phys Chem A 2001, 105, 9736.
34. Hobza, P.; Selzle, H. L.; Schlag, E. W. J Phys Chem 1996, 100, 18790.
35. Frenking, G.; Antes, I.; Böhme, M.; Dapprich, S.; Ehlers, A. W.; Jonas, V.; Neuhaus, A.; Otto, M.; Stegmann, R.; Veldkamp, A.; Vyboishchikov, S. F. Reviews in Computational Chemistry, Vol. 8; VCH: New York, 1996.
36. Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. X. J Phys Chem A 1999, 103, 1913.
37. Pulay, P.; Saebø, S.; Wolinski, K. Chem Phys Lett 2001, 344, 543.
38. Ayala, P. Y.; Scuseria, G. E. J Chem Phys 1999, 110, 3660.

39. Ayala, P. Y.; Kudin, K. N.; Scuseria, G. E. J Chem Phys 2001, 115, 9698.
40. Jung, Y.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. J Chem Phys 2004, 121, 9793.
41. Lochan, R. C.; Jung, Y.; Head-Gordon, M. J Phys Chem A 2005, 109, 7598.
42. Grimme, S. J Chem Phys 2003, 118, 9095.
43. Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Chem Phys Lett 1993, 208, 359.
44. Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. Chem Phys Lett 1998, 294, 143.
45. Saebø, S.; Pulay, P. Ann Rev Phys Chem 1993, 44, 213.
46. Maslen, P. E.; Head-Gordon, M. Chem Phys Lett 1998, 283, 102.
47. Maslen, P. E.; Head-Gordon, M.; J Chem Phys 1998, 109, 7093.
48. Lee, M. S.; Maslen, P. E.; Head-Gordon, M. J Chem Phys 2000, 112, 3592.
49. Schütz, M.; Hetzer, G.; Werner, H. J. J Chem Phys 1999, 111, 5691.
50. Subotnik, J. E.; Head-Gordon, M. J Chem Phys 2005, 122, 034109.
51. Werner, H. J.; Manby, F. R.; Knowles, P. J. J Chem Phys 2003, 118, 8149.
52. DiStasio, R. A.; Jung, Y. S.; Head-Gordon, M. J Chem Theory Comput 2005, 1, 862.
53. El Azhary, A.; Rauhut, G.; Pulay, P.; Werner, H. J. J Chem Phys 1998, 108, 5185.
54. Schütz, M.; Werner, H. J.; Lindh, R.; Manby, F. R. J Chem Phys 2004, 121, 737.
55. Russ, N.; Crawford, T. D. J Chem Phys 2004, 121, 691.
56. Weigend, F.; Häser, M. Theor Chem Acc 1997, 97, 331.
57. Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Chem Phys Lett 1990, 166, 275.
58. Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Chem Phys Lett 1990, 166, 281.
59. Haase, F.; Ahlrichs, R. J Comput Chem 1993, 14, 907.
60. Nielsen, I. M. B. Chem Phys Lett 1996, 255, 210.
61. Head-Gordon, M. Mol Phys 1999, 96, 673.
62. Hättig, C.; Weigend, F. J Chem Phys 2000, 113, 5154.
63. Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. Theor Chem Acc 2003, 110, 233.
64. Handy, N. C.; Schaefer, H. F. J Chem Phys 1984, 81, 5031.
65. Shao, Y.; Fusti-Molnar, L.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F.; Kong, J.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Phys Chem Chem Phys 2006, 8, 3172.
66. Becke, A. D. J Chem Phys 1993, 98, 5648.
67. Lee, C.; Yang, W.; Parr, R. G. Phys Rev B 1998, 37, 785.
68. Dunning, T. H. J Chem Phys 1989, 90, 1007.
69. Schäfer, A.; Horn, H.; Ahlrichs, R. J Chem Phys 1992, 97, 2571.
70. Schäfer, A.; Huber, C.; Ahlrichs, R. J Chem Phys 1994, 100, 5829.
71. Hariharan P. C.; Pople J. A. Theoret Chimica Acta 1973, 28, 213.
72. Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. J Chem Phys 1980, 72, 650.
73. Francl, M. M.; Petro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. J Chem Phys 1982, 77, 3654.
74. Blaudeau, J. P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. J Chem Phys 1997, 107, 5016.
75. Rassolov, V.; Pople, J. A.; Ratner, M.; Windus T. L. J Chem Phys 1998, 109, 1223.
76. Weigend, F.; Kohn, A.; Hättig, C. J Chem Phys 2002, 116, 3175.
77. Bernholdt, D. E.; Harrison, R. J. J Chem Phys 1998, 109, 1593.
78. Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. J Chem Phys 1991, 94, 7221.
79. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. J Chem Phys 1998, 109, 7764.
80. Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. J Chem Phys 2000, 112, 7374.
81. Herbert, J. M.; Head-Gordon, M. Phys Chem Chem Phys 2005, 18, 3269.
82. (a) Huber, K. P.; Herzberg, G.;Constants of Diatomic Molecules, Molecular Structure and Molecular Spectra, Vol. 4; Van Nostrand Reinhold: Princeton, 1979.; (b) Laurie, V. W.; Kuczkowski, R. L.; Schwendeman, R. H.; Ramsay, D. A.; Lovas, F. J.; Lafferty, W. J.; Maki, A. G. J Phys Chem Ref Data 1979, 8, 619.; (c) Graner, G.; Kuchitsu, K.; Structure of Free Polyatomic Molecules: Basic Data; Springer: Place, 1998.
83. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. J Am Chem Soc 1997, 119, 5908.
84. Rhee, Y. M. J Chem Phys 2000, 113, 6021.
85. Spartan '04 for Windows; Wavefunction, Inc.: Irvine, 2004.
86. Jeffrey, G. A. An Introduction to Hydrogen Bonding; Oxford: New York, 1997.
87. Yu, C.-H.; Norman, M. A.; Schäfer, L.; Ramek, M.; Peeters, A.; van Alsenoy, C. J Mol Struct 2001, 567/568, 361.
88. Chaudhuri, P.; Canuto, S. J Mol Struct (Theochem) 2002, 577, 267.
89. Polly, R.; Werner, H. J.; Manby, F. R.; Knowles, P. J. Mol Phys 2004, 102, 2311.
90. Liang, W.; Head-Gordon, M. J Phys Chem A 2004, 108, 3206.
91. Steele, R. P.; Shao, Y.; DiStasio, R. A.; Kong, J.; Head-Gordon, M. J Chem Phys 2006, 125, 074108.