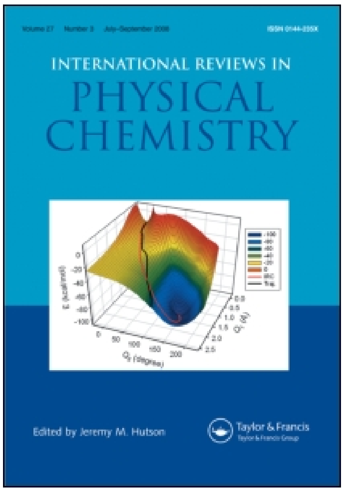

International Reviews in Physical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713724383

A review of canonical transformation theory

Eric Neuscamman[a]; Takeshi Yanai[b]; Garnet Kin-Lic Chan[a]
[a] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA [b] Institute of Molecular Science, Myodaiji, Okazaki, Japan

To cite this Article Neuscamman, Eric , Yanai, Takeshi and Chan, Garnet Kin-Lic(2010) 'A review of canonical transformation theory', International Reviews in Physical Chemistry, 29: 2, 231 — 271
To link to this Article: DOI: 10.1080/01442351003620540
URL: http://dx.doi.org/10.1080/01442351003620540

International Reviews in Physical Chemistry
Vol. 29, No. 2, April–June 2010, 231–271

Taylor & Francis
Taylor & Francis Group

A review of canonical transformation theory

Eric Neuscamman[a], Takeshi Yanai[b] and Garnet Kin-Lic Chan[a*]

[a]*Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA;* [b]*Institute of Molecular Science, 38 Nishigo-Naka, Myodaiji, Okazaki 444-8585, Japan*

(*Received 29 September 2009; final version received 13 January 2010*)

Canonical transformation (CT) theory targets the description of dynamic correlation in multireference quantum chemistry problems. When combined with a static correlation quantum chemistry method, it enables the quantitative description of chemical processes involving electronic structure not described by a single electronic configuration. We argue that many multireference dynamic correlation methods display unsatisfactory characteristics, including lack of size-consistency, a low-order treatment of correlation, and a poor computational scaling. By contrast, CT theory is based on an exponential ansatz that is rigorously size-consistent, reduces in a single-reference limit to a coupled cluster theory, and has an n^6 computational scaling with system and active space size. The efficient formulation of CT theory has allowed it to be applied to difficult systems in conjunction with active spaces with more than 30 orbitals, beyond the reach of traditional methods, with an accuracy that far exceeds multireference perturbation theories. Here we review the basic motivation, formulation, and implementation of CT theory, as well as survey some of our recent applications and possible future directions.

Keywords: correlation methods; canonical transformation theory; density matrix renormalization group; coupled cluster theory; configuration interaction; perturbation theory

Contents

---

*Corresponding author. Email: gc238@cornell.edu

ISSN 0144–235X print/ISSN 1366–591X online

DOI: 10.1080/01442351003620540
http://www.informaworld.com

1. Introduction

1.1. *Overview*

This review describes our work on the canonical transformation theory of electron correlation, as it appears in several journal articles and preprints [1–6]. The review is divided into several sections. Section 1 provides an overview of the dynamic correlation problem in the multireference quantum setting, why canonical transformation theory was developed, and its relation to earlier theories. Section 2 provides the theory's mathematical foundation and a discussion of its formal properties. Section 3 gives a formal description of how the ansatz is optimized, while Section 4 presents the details of our computational algorithm. Section 5 is an overview of a number of interesting special topics. Section 6 provides a survey of some of the quantum chemical problems that have been studied using the canonical transformation theory. Finally, Section 7 describes some of the remaining challenges, possible future research directions, and our conclusions.

1.2. *The challenge of dynamic correlation in multireference quantum chemistry*

In this review, we consider the challenge of describing dynamic correlation in multireference systems. Before proceeding further, it is useful to define multireference, as well as the distinction between static and dynamic electron correlation. In most molecules, atoms arrange themselves so as to maximize the bonding overlap between their valence orbitals, which typically creates an energy gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO, respectively). When this gap is large, the molecule's wavefunction is dominated by a single electronic configuration, and we say that it is a 'single-reference' system. The valence electrons behave as non-interacting fermions moving in a mean-field potential and can be described correctly (at least qualitatively) using Hartree–Fock theory. When the HOMO/LUMO gap is small, however, there can be many low-energy electronic configurations, producing a wavefunction that is a superposition of multiple configurations without any one being dominant. These molecules represent 'multireference' systems, for which a mean-field description of the electrons is often qualitatively incorrect. Instead, the complete active space self-consistent field (CASSCF) description should be employed, in which the wavefunction is explicitly constructed as a superposition of all possible configurations of the valence electrons among the valence orbitals [7,8]. Unlike Hartree–Fock theory, CASSCF can describe strong correlations between the valence electrons. This 'static' or 'non-dynamic' correlation differentiates multireference systems from single-reference systems and can be found in a number of chemical settings, including:

- Chemical reactions. Reaction pathways often contain points at which the valence configurations of the reactants and products are nearly degenerate.
- Excited states. Away from the ground-state, different sets of excitations can yield configurations of similar energy. Consequently, many excited states are found to be mixtures of several configurations.
- Poor orbital overlap. When bonding orbitals do not overlap well at chemical distances, the bonding/antibonding energy splitting is small and multiple configurations may be nearly degenerate. This phenomenon is common with transition metal d orbitals.

In all of these examples, a CASSCF or similar wavefunction is required to create a superposition of valence configurations in order to capture the static correlation and yield a qualitatively correct description of the chemistry.

Although a pure valence description of the electronic structure is conceptually appealing, it is well known that capturing the static correlation alone is not sufficient to achieve a quantitatively accurate chemical description. Here we take chemically accurate to mean about 1 kcal mol^{-1} for reaction energies, and about 0.1 eV for excitation energies. Instead, the electronic structure must be described in a much larger basis which includes both higher angular momentum non-valence atomic orbitals as well as diffuse Rydberg-like functions. We can understand this requirement in several ways. Firstly, the shapes of orbitals should depend on their occupancies: in a configuration where a valence orbital is doubly occupied, we would expect electrons to sit in somewhat different orbits than in the case of single occupancy, reflecting radial and angular correlations. Also, Coulomb repulsion can lead to a small probability of electronic configurations with occupancy of non-valence orbitals, so that electrons can better avoid each other.

These kinds of adjustments to the valence electronic structure constitute ‘dynamic’ correlation, which is not captured by either the Hartree–Fock or CASSCF wavefunctions. The quantitative description of electronic structure therefore requires additional theoretical models for dynamic correlation, which can be separated into two groups: those for single-reference systems and those for multireference systems. As we shall see, multireference dynamic correlation theories have so far been less successful than their single-reference counterparts.

When the valence electronic structure is qualitatively captured by a single electronic configuration, the description of dynamic correlation is well understood. We shall refer to this limit as the single-reference dynamic correlation problem. Within a wavefunction setting, three common approaches include:

- Møller–Plesset perturbation theory, a Rayleigh–Schrödinger perturbation theory where the Fock operator is the zeroth order Hamiltonian [9]. The perturbation series is usually truncated at second-order (MP2). MP2 calculations have a formal computational scaling of n^5 with the molecular size n.
- Configuration interaction, a variational ansatz formed by a linear combination of configurations including excitations from the reference up to a given excitation rank, typically singles and doubles replacements (CISD) [9]. CISD is not size consistent and is not usually used without a size consistency correction. Approximate size consistent CI theories can be obtained in various ways, such as through the Davidson correction [10,11], or through the coupled-pair-functional [12]. CISD has a formal computational scaling of n^6 with molecular size.
- Coupled cluster theory, which uses an exponential form for the excitation operator [9,13]. Formally, coupled cluster theory sums many terms in perturbation theory to infinite order, and most importantly all those terms which are necessary for exact size consistency. The most common coupled cluster theory uses singles and doubles excitation operators (CCSD) and has an n^6 computational scaling with molecular size.

Out of these three approaches for single-reference dynamic correlation, coupled cluster theory has established itself as the most satisfactory, both formally and in terms of numerical performance. It is sometimes referred to as the ‘gold-standard’ of dynamic correlation methods [13].

In this review, however, we are concerned with those chemical situations where the valence electronic structure requires a *multireference* wavefunction description. We shall refer to the dynamic correlation in this limit as multireference dynamic correlation. The process of including dynamic correlation corrections is now made much more complicated as one cannot use the many simplifications that arise from a single-reference starting point. Nonetheless, multireference analogues of the single-reference dynamic correlation theories have been considered, including:

- Multireference perturbation theory, such as the complete active space perturbation theory (CASPT) [14], multireference Møller–Plesset (MRMP) [15], and n-electron valence perturbation theory (NEVPT) [16,17]. These constitute the most widely applied multireference dynamic correlation theories. Unlike in the single-reference limit, some formulations of multireference perturbation theory are not exactly size consistent. The computational scaling of multireference perturbation theory depends on the specific formulation, but is generally higher than Møller–Plesset theory, with a scaling of at least n^7 for the fully internally contracted variant.

- Multireference configuration interaction (MRCI) [18–20] with approximate size consistency corrections (MRCI+Q) and averaged coupled pair functionals (MRACPF, MRAQCC) [21,22]. These are divided into two types: internally contracted (the variational space is formed from excitation operators acting on a single starting multireference wavefunction) or externally contracted (the variational space is formed by excitations from all determinants in the active space). The computational scalings are formulation dependent but are typically n^{10} and e^n, respectively.
- Multireference coupled cluster theories (MRCC) [23]. There are many variants of MRCC which reflects the mathematical difficulties in extending the CC formalism beyond a single-reference starting point. In general MRCC theories have computational costs that are even higher than MRCI theory. While there is much activity in this field, MRCC methods have yet to be applied to realistic problems.

From this brief analysis, it is clear that the description of dynamic correlation with multireference wavefunctions is much less satisfactory than in the single-reference case. Because of this handicap, in certain situations using a multireference (e.g. complete active space) description of the valence electrons – in principle a more flexible theoretical framework – can lead to a poorer quantitative accuracy than a single-reference description. Consider, for example, a chemical transformation where the electronic structure changes from single-reference to multireference in character. A typical multireference description, e.g. through CASPT2, obtains the static correlation of the active valence orbitals exactly (by virtue of the CASSCF treatment) but recovers only a small portion of the dynamic correlation of the external orbitals. This is appropriate when the pure-valence static correlation is larger than the external contributions, but in the single-reference limit all the correlation is dynamic, and the external contributions may in fact be larger, given that there is no degeneracy in the valence space and there are many more external orbitals. A single-reference coupled cluster description provides a (perturbatively) high-order description of dynamic correlation for *all* degrees of freedom. It is likely that in the single-reference limit of the chemical transformation, the coupled cluster treatment would capture a larger portion of the correlation energy than the CASPT2 description, unless a very large active space were to be used. We see that CASPT2 is, in essence, biased against the single-reference limit, due to the inadequate treatment of dynamic correlation. (In Section 6 we see a real-life example in the $[Cu_2O_2]^{2+}$ isomerization curve.) Even when such behavior is not of concern, multireference dynamic correlation theories are much more costly than their single-reference counterparts, limiting their application to very small systems. These various unsatisfactory aspects of existing multireference dynamic correlation theories clearly motivate the development of a new theoretical model.

1.3. *Desirable features for a dynamic correlation theory*

Before considering a new multireference dynamic correlation theory, we should first ask, what are the desirable features? In our view, a multireference dynamic correlation theory should

- be size consistent (to allow meaningful calculations on large systems),
- not bias the chemistry towards multireference electronic structure by underestimating the dynamic correlation in single-reference settings,
- reduce in single-reference settings to a coupled-cluster like theory,
- have a reasonable scaling with system size, such as CCSD's n^6 scaling.

The canonical transformation theory of dynamical correlation is designed to obtain these features. In particular, it

- utilizes an exponential excitation operator, which preserves size consistency,
- provides a coupled cluster-like description of dynamic correlation; in single-reference settings, this guarantees an accurate description of the correlation,
- has a favorable computational scaling of n^6 in the system size that is achieved through the use of operator and cumulant decompositions.

1.4. *Connections to earlier work*

Canonical transformation theory derives from several earlier theoretical developments. The main ingredients of canonical transformation theory are (see Section 2 for more details) (i) a unitary exponential form for the dynamic correlation, (ii) operator and cumulant decompositions to simplify the energy and amplitude equations, and (iii) an emphasis on an effective Hamiltonian picture of the dynamic correlation. Our initial motivation came from White's work on numerical canonical transformations [24]. When rewritten in an appropriate form this can be seen to contain both (i) and (iii) above, but without the systematic simplifications presented by operator and cumulant decompositions. However, there are many other developments in quantum chemistry with a direct connection to canonical transformation theory, including

- Coupled cluster theory. While the exponential operator in ordinary CC theory is not unitary, unitary variants of coupled cluster theory were explored in a single-reference setting by Kutzelnigg [25,26], Bartlett [27–29], and Pal [30,31]. Multireference unitary coupled cluster theory was developed by Simons [32–34].
- Operator decompositions and the generalized Wick's theorem, introduced by Mukherjee and Kutzelnigg [35–37].
- Density matrix and cumulant decompositions [37–41]. While their most widespread use in conjunction with the density equation, also known as the contracted Schrödinger equation [38–40], is quite different from the use in canonical transformation theory, there is a close connection to Kutzelnigg and Mukherjee's irreducible contracted Schrödinger equation theory [42] and the related anti-hermitian contracted Schrödinger equation theory introduced by Mazziotti [43].
- There is much earlier work on effective Hamiltonians, including Freed's effective valence Hamiltonian theory [44], Kirtman and Hoffmann's generalized van Vleck

theories [45,46], and the equation-of-motion coupled cluster theory and symmetry-adapted cluster configuration interaction theory [47,48].

2. Theory

2.1. *Basics*

CT theory seeks to find the unitary transformation that maps a reference wavefunction $|\Psi_0\rangle$ onto the true wavefunction $|\Psi\rangle$, as shown in Equation (1).

$$|\Psi\rangle = e^{A}|\Psi_0\rangle. \tag{1}$$

The reference function is usually assumed to contain the correct static correlation for the problem and thus the transformation operator e^{A} is thought of as introducing the dynamic correlation. The transformation in Equation (1) is of the coupled cluster form, but unlike in CC theory, CT theory is constructed in such a way that the only information we require from the reference function is its one- and two-body RDMs. This allows CT theory to efficiently treat dynamic correlation not only for the Hartree–Fock reference but also for multi-configurational references, such as GVB, CASSCF, and DMRG-SCF wavefunctions. In singles and doubles CT, the unitary transformation is built from the exponential of some combination of anti-symmetric single and double excitation operators $\hat{o}$,

$$A = -A^{\dagger} = \sum_i C_i \hat{o}_i. \tag{2}$$

These operators can in principle be any set of one- and two-body excitation operators $(a_p^{\dagger}a_q - a_q^{\dagger}a_p)$ and $(a_p^{\dagger}a_q^{\dagger}a_s a_r - a_r^{\dagger}a_s^{\dagger}a_q a_p)$. However, if we assume that the reference $|\Psi_0\rangle$ correctly describes the static correlation in some complete active space, it is natural to consider only those excitations which change the occupancies of at least one external (non-active) orbital. Denoting core orbitals by c, active by a, and virtual by v, and using the notation $a_q^p = a_p^{\dagger}a_q$ and $a_{rs}^{pq} = a_p^{\dagger}a_q^{\dagger}a_s a_r$, the full set of possible singles and doubles excitations are given by the following, in which all indices are summed over.

$$\begin{aligned}
A = {}& A_{c_1c_2}^{a_1a_2}(a_{c_1c_2}^{a_1a_2} - a_{a_1a_2}^{c_1c_2}) + A_{a_3c_1}^{a_1a_2}(a_{a_3c_1}^{a_1a_2} - a_{a_1a_2}^{a_3c_1}) + A_{c_1}^{a_1}(a_{c_1}^{a_1} - a_{a_1}^{c_1}) && \text{core-active} \\
& + A_{a_1a_2}^{v_1v_2}(a_{a_1a_2}^{v_1v_2} - a_{v_1v_2}^{a_1a_2}) + A_{a_1a_2}^{a_3v_1}(a_{a_1a_2}^{a_3v_1} - a_{a_3v_1}^{a_1a_2}) + A_{a_1}^{v_1}(a_{a_1}^{v_1} - a_{v_1}^{a_1}) && \text{virtual-active} \\
& + A_{a_1c_1}^{a_2v_1}(a_{a_1c_1}^{a_2v_1} - a_{a_2v_1}^{a_1c_1}) + A_{a_1c_1}^{v_1v_2}(a_{a_1c_1}^{v_1v_2} - a_{v_1v_2}^{a_1c_1}) + A_{c_1c_2}^{a_1v_1}(a_{c_1c_2}^{a_1v_1} - a_{a_1v_1}^{c_1c_2}) && \text{core-virtual-active} \\
& + A_{c_1c_2}^{v_1v_2}(a_{c_1c_2}^{v_1v_2} - a_{v_1v_2}^{c_1c_2}) + A_{c_1}^{v_1}(a_{c_1}^{v_1} - a_{v_1}^{c_1}) && \text{core-virtual.}
\end{aligned} \tag{3}$$

Note that each excitation operator (e.g. $a_{a_1}^{v_1}$) must be accompanied by a de-excitation operator $(a_{v_1}^{a_1})$ in order to make CT theory unitary. We will typically refer to the combined operators $(a_{a_1}^{v_1} - a_{v_1}^{a_1})$ as excitation operators, but the reader should remember that they are really anti-symmetrized excitation operators that contain a de-excitation component. In practice, further restrictions of the operators are used to simplify their optimization (see Section 3).

Instead of solving the electronic Schrödinger equation in the direct form,

$$He^{A}|\Psi_0\rangle = Ee^{A}|\Psi_0\rangle, \tag{4}$$

CT theory works with a quasi-particle Schrödinger equation with an effective Hamiltonian,

$$e^{-A}He^{A}|\Psi_0\rangle = \bar{H}|\Psi_0\rangle = E|\Psi_0\rangle. \tag{5}$$

Equation (5) is interpreted as saying that there exists a reference function of quasi-particles, interacting via the effective quasi-particle Hamiltonian $\bar{H}$ which has the same energy as the original system of electrons interacting through the bare Hamiltonian H. This canonical (unitary) transformation from electrons to quasi-particles moves the complexity of the correlation problem from the wavefunction to the Hamiltonian. For electrons, the known Hamiltonian has a number of integrals proportional to the fourth power of the number of electrons, while the exact wavefunction (which must be determined) has a complexity in its determinantal expansion that is exponential in the electron number. For quasi-particles, the exact wavefunction $|\Psi_0\rangle$ is the comparably simple reference wavefunction (which is known) while the unknown effective Hamiltonian has exponential complexity. This complexity of the effective Hamiltonian is revealed by the Baker–Campbell–Hausdorff (BCH) expansion,

$$\bar{H} = H + [H, A] + \frac{1}{2!}[[H, A], A] + \cdots, \tag{6}$$

where each commutator creates a successively higher-particle rank operator. We therefore see that the effective Hamiltonian can in general contain interactions between any number of quasi-particles (up to the maximum number present in the system). In contrast, electrons only interact pairwise in the bare Hamiltonian.

So far in our development, the canonical transformation has only accomplished a transfer of complexity from the wavefunction to the Hamiltonian. Where this conversion becomes useful is in devising approximations based on our intuition about electrons and quasi-particles. For a sufficiently accurate reference wavefunction, A will be small and electrons and quasi-particles will have similar physics. We can therefore make an approximation that the quasi-particles, like electrons, have no direct interactions more complex than pairwise interactions, which ensures that the effective Hamiltonian $\bar{H}$ has the same quartic complexity as the original Hamiltonian. This is the defining approximation of CT theory.

Naturally, we want to retain the effects of the higher than pairwise interactions in some average way. We achieve this in the effective Hamiltonian by approximating three- and higher-body interactions (operators) with products of one- and two-body interactions, a procedure we refer to as operator decomposition. The decomposition is accomplished by replacing the commutators that appear in the definition of $\bar{H}$, Equation (6), with approximations $[H, A] \rightarrow [H, A]_{1,2}$, in which the three-body operators resulting from the commutator are approximated by one- and two-body operators. The effective Hamiltonian may now be constructed with only a polynomial (n^6) amount of effort through the recursive application of this commutator approximation:

$$\bar{H} = H + [H, A]_{1,2} + \frac{1}{2!}[[H, A]_{1,2}, A]_{1,2} + \cdots \tag{7}$$

How do we obtain the optimal canonical transformation to $\bar{H}$? To answer this question, we return to the quasi-particle Schrödinger equation. Recalling that A is parametrized in terms of a set of coefficients C_i, we require that these satisfy Equation (5) in a projective sense with respect to the first-order interacting space $\{\hat{o}_i|\Psi_0\rangle\}$. This yields the coupled cluster style amplitude equations $\langle\Psi_0|\hat{o}_i^\dagger\bar{H}|\Psi_0\rangle = 0$, which we rearrange to form the CT amplitude equations,

$$\langle\Psi_0|[\bar{H}, \hat{o}_i]_{1,2}|\Psi_0\rangle = 0, \tag{8}$$

which take the form of generalized Brillouin conditions [49]. As with the BCH expansion, these amplitude equations employ the commutator approximation so that they may be evaluated in n^6 time. In practice, solving these equations is not trivial and requires additional restrictions on the operators $\hat{o}_i$, which we discuss in detail in Section 3.

Once the amplitude equations have been solved, the amplitude operator A is used to construct the effective Hamiltonian using Equation (7). The CT energy is then defined by tracing the effective integrals with the reference function's RDMs,

$$E = \langle\Psi_0|\bar{H}|\Psi_0\rangle. \tag{9}$$

To summarize, Equations (7)–(9) constitute the working equations of CT theory. Through the use of the commutator approximation, the CT energy and effective Hamiltonian may be evaluated at an n^6 cost using only the reference function's one- and two-body RDMs and the one- and two-body operators of the electronic Hamiltonian.

2.2. *Operator decomposition*

As discussed above, the central approximation in CT theory is to replace the commutator $[H, A]$, which when evaluated exactly contains three-body operators, with an approximation $[H, A]_{1,2}$ containing only one- and two-body operators. Using Mukherjee and Kutzelnigg's formalism of extended normal ordering (ENO) [35–37], it is possible to construct an approximation for a three-body operator that has the same expectation value as the original operator. Just as a traditionally normal ordered operator has a zero expectation value with respect to the true vacuum, an ENO operator has a zero expectation value with respect to a general (possibly multi-configurational) reference state. In extended normal ordered theory, a three-body operator may be written as the sum of an ENO three-body operator, a combination of ENO one- and two-body operators, and the expectation value of the original operator in the form of a three-body RDM element. This can be seen through the following definitions of the one-, two-, and three-body ENO operators $\tilde{a}^{p_1}_{q_1}$, $\tilde{a}^{p_1p_2}_{q_1q_2}$, $\tilde{a}^{p_1p_2p_3}_{q_1q_2q_3}$, in which γ represents the reference function's RDMs, e.g. $\gamma^p_q = \langle\Psi_0|a^p_q|\Psi_0\rangle$.

$$\tilde{a}^{p_1}_{q_1} = a^{p_1}_{q_1} - \gamma^{p_1}_{q_1} \tag{10a}$$

$$\tilde{a}^{p_1p_2}_{q_1q_2} = a^{p_1p_2}_{q_1q_2} - \sum(-1)^x\gamma^{p_1}_{q_1}\tilde{a}^{p_2}_{q_2} - \gamma^{p_1p_2}_{q_1q_2} \tag{10b}$$

$$\tilde{a}^{p_1p_2p_3}_{q_1q_2q_3} = a^{p_1p_2p_3}_{q_1q_2q_3} - \sum(-1)^x\gamma^{p_1}_{q_1}\tilde{a}^{p_2p_3}_{q_2q_3} - \sum(-1)^x\gamma^{p_1p_2}_{q_1q_2}\tilde{a}^{p_3}_{q_3} - \gamma^{p_1p_2p_3}_{q_1q_2q_3}. \tag{10c}$$

These equations make use of the notation $\sum(-1)^x A^{p_1 p_2 \cdots}_{q_1 q_2 \cdots} B^{p_k p_{k+1} \cdots}_{q_k q_{k+1} \cdots} \cdots$, which implies that there is one term for each unique partitioning of the indices among the objects $(A, B, \ldots)$ in which p_i are kept on top and q_i on bottom. For each permutation of the indices from their original positions, a factor of (-1) is applied. We recognize that the three-body ENO operator represents three-body fluctuations from the operator's average one- and two-particle like behavior, represented by the RDM element and the one- and two-body ENO operators. If the final wavefunction is close to the reference function (which is our working assumption in CT theory), we hope that the effect of neglecting these three-body fluctuations is small as their expectation value with respect to the reference is zero. Therefore, in forming our commutator approximation, we convert all the operators produced by the commutator to ENO form and neglect any resulting three-body ENO operators,

$$a^{p_1 p_2 p_3}_{q_1 q_2 q_3} \Rightarrow \sum(-1)^x \gamma^{p_1}_{q_1} \tilde{a}^{p_2 p_3}_{q_2 q_3} + \sum(-1)^x \gamma^{p_1 p_2}_{q_1 q_2} \tilde{a}^{p_3}_{q_3} + \gamma^{p_1 p_2 p_3}_{q_1 q_2 q_3}. \tag{11}$$

Now only constants and one- and two-body operators remain, as originally desired. This allows us to evaluate the BCH expansion recursively, as each term's approximation may be used to approximate the inner commutators of the next term. If we let $\bar{H}^{(n)}$ represent the approximation to the nth term in the BCH expansion, then this recursive evaluation leads to the formula $\bar{H}^{(n)} = \frac{1}{n}[\bar{H}^{(n-1)}, A]_{1,2}$. The total effective Hamiltonian is then approximated as

$$\bar{H} = e^{-A} H e^{A} \simeq \sum_{n=0}^{\infty} \bar{H}^{(n)}, \tag{12}$$

with $\bar{H}^{(0)}$ equal to the unmodified electronic Hamiltonian H. Note that while this sum is formally infinite, the amplitude operator A is assumed to be small and in practice the terms decay quickly enough that only a finite number need to be evaluated (typically 8–10 for a precision of 10^{-9} Hartrees).

2.3. *Cumulants*

So far, our commutator approximation requires the use of the reference function's three-body RDM. In practice, however, the three-body RDM is not always readily available, and thus we would like to limit ourselves to using only the reference's one- and two-body RDMs. To accomplish this goal we turn to the theory of density matrix cumulants [38,39,41], which represent the irreducible correlations of a given particle rank present in the RDMs. If n-body interactions are unimportant, then neglecting the n-body cumulant creates a reasonable approximation for the n-body RDM consisting of antisymmetric products of lower particle rank RDMs. This approximation technique has been most widely used in contracted Schrödinger equation (CSE) theories [38,40,43,50,51] where the one- and two-body RDMs are optimized directly and the three- and possibly four-body RDMs are approximated by neglecting the three- and four-body cumulants. In CT theory's commutator approximation we similarly neglect the three-body cumulant, reducing the three-body RDM that arises from operator decomposition to a collection of antisymmetric products of one- and two-body RDMs.

Therefore, the approximation of the three-body operator shown in Equation (11) becomes [4]

$$\begin{aligned}(a^{p_1p_2p_3}_{q_1q_2q_3})_{1,2} &= \sum(-1)^x \gamma^{p_1}_{q_1} a^{p_2p_3}_{q_2q_3} && \text{(9 terms)}\\ &+\sum(-1)^x \gamma^{p_1p_2}_{q_1q_2} a^{p_3}_{q_3} && \text{(9 terms)}\\ &-2\sum(-1)^x \gamma^{p_1}_{q_1}\gamma^{p_2}_{q_2} a^{p_3}_{q_3} && \text{(18 terms)}\\ &-\sum(-1)^x \gamma^{p_1}_{q_1}\gamma^{p_2p_3}_{q_2q_3} && \text{(9 terms)}\\ &+4\sum(-1)^x \gamma^{p_1}_{q_1}\gamma^{p_2}_{q_2}\gamma^{p_3}_{q_3} && \text{(6 terms).}\end{aligned} \tag{13}$$

The result is a commutator approximation $[H, A]_{1,2}$ that contains only one- and two-body operators and requires knowing only the one- and two-body RDMs of the reference wavefunction. The analogous approximation for spin free operators can be arrived at in the same manner by neglecting the spin free three-body ENO operator and cumulant.

While the use of cumulants in CT theory is similar to that in CSE theory, there are some important differences that help CT theory avoid common pitfalls associated with CSE methods. One such issue is the dependence of N-representability error in the approximated three-body RDM on basis set size, as explored by Harris [52] and Herbert [53]. As opposed to CSE methods, which approximate the final wavefunction's three-body RDM in the entire orbital space, CT theory approximates only the reference wavefunction's three-body RDM in the active space. This greatly reduces the problem of basis set dependence, as the reference wavefunction's active space RDMs change little with the basis set. Indeed, our previous results [2] show that CT theory's accuracy is not affected when changing from a double- to triple-zeta basis set in the nitrogen dimer. An additional difference from CSE methods is that CT theory optimizes an excitation operator (as in CC theory) that can modify the active space correlation only indirectly via its coupling to the external space, rather than the active space RDMs directly, as is done in CSE. It is therefore reasonable to expect N-representability problems to be less prevalent in CT theory, and indeed we find this to be the case in practice.

2.4. *Size consistency*

CT theory is rigorously size consistent, which we shall now demonstrate. Consider two non-interacting systems X and Y, for which the electronic Hamiltonian can be written as $H = H_X + H_Y$. Treating the systems separately, CT theory will produce two amplitude solutions A_X and A_Y as well as two effective Hamiltonians $\bar{H}_X = e^{-A_X} H_X e^{A_X}$ and $\bar{H}_Y = e^{-A_Y} H_Y e^{A_Y}$. These solutions satisfy the amplitude equations for the systems individually:

$$\langle \Psi_{X_0} | [\bar{H}_X, \hat{o}_i]_{1,2} | \Psi_{X_0} \rangle = 0, \tag{14}$$

$$\langle \Psi_{Y_0} | [\bar{H}_Y, \hat{o}_i]_{1,2} | \Psi_{Y_0} \rangle = 0. \tag{15}$$

The total energy for the two systems is simply the sum of the expectation values of their effective Hamiltonians,

$$E = \langle \Psi_{X_0} | \bar{H}_X | \Psi_{X_0} \rangle + \langle \Psi_{Y_0} | \bar{H}_Y | \Psi_{Y_0} \rangle = E_X + E_Y. \tag{16}$$

If we instead treat the systems together, choose $A = A_X + A_Y$, and make use of the fact that operators from different systems commute, we find that the effective Hamiltonian is the sum of the separate systems' effective Hamiltonians,

$$\bar{H} = e^{-(A_X+A_Y)} H e^{(A_X+A_Y)} = e^{-A_X} H_X e^{A_X} + e^{-A_Y} H_Y e^{A_Y} = \bar{H}_X + \bar{H}_Y. \tag{17}$$

This separation allows the amplitude equation,

$$\langle \Psi_{X_0} | \langle \Psi_{Y_0} | [\bar{H}_X + \bar{H}_Y, \hat{o}_i]_{1,2} | \Psi_{X_0} \rangle | \Psi_{Y_0} \rangle = 0, \tag{18}$$

to be satisfied for the two possible types of excitation operators $\hat{o}_i$. If $\hat{o}_i$ operates entirely in system X or entirely in system Y, then Equation (18) is satisfied due to Equations (14)–(15) and the fact that $\hat{o}_i$ commutes with the effective Hamiltonian of the other system. If $\hat{o}_i$ operates on both systems X and Y simultaneously, it either changes the particle number in each system or simultaneously excites into both systems' external spaces. In this case Equation (18) is satisfied because the effective Hamiltonian $\bar{H}_X + \bar{H}_Y$ cannot transfer particles between the systems and cannot simultaneously de-excite from both systems' external spaces. Thus we see that the solution A to the CT equations for the combined system is simply the sum of the solutions A_X and A_Y of the separate systems. Finally, by taking the expectation value of the combined system's effective Hamiltonian, we see that the energy is the same as for the separate systems and that CT theory is rigorously size consistent.

$$\begin{aligned}\langle \Psi_0 | \bar{H} | \Psi_0 \rangle &= \langle \Psi_{Y_0} | \langle \Psi_{X_0} | (\bar{H}_X + \bar{H}_Y) | \Psi_{X_0} \rangle | \Psi_{Y_0} \rangle \\ &= \langle \Psi_{X_0} | \bar{H}_X | \Psi_{X_0} \rangle + \langle \Psi_{Y_0} | \bar{H}_Y | \Psi_{Y_0} \rangle \\ &= E_X + E_Y. \end{aligned} \tag{19}$$

2.5. *Perturbative analysis*

In this section we analyze the CT energy perturbatively first for the case in which the reference function is a single Slater determinant, and later for the general case of a multi-configurational reference function. This analysis is instructive as it provides insight into the similarities and differences between CT theory and Coupled Cluster (CC) theory. This section follows the analysis of Bartlett *et al.* in Refs. [27,28,54]. For the purposes of this section, we separate the CT amplitude operator into its one- and two-body components via $A = A_1 + A_2$.

First consider a Hartree–Fock reference function and transform to the Fermi vacuum (all occupied orbitals are in the vacuum). Then all Fermi-vacuum particle density matrices are zero and the commutator approximation corresponds to simply neglecting all three-body operators. This type of operator truncation is used in the canonical diagonalization theory of White [24].

Now write the Hamiltonian as $H = E_{\text{HF}} + F + W$, where F is the one-body Fock operator and W is the two-body fluctuation potential. From Brillouin's theorem, we recognize that A_2 is first order in W, while A_1 is second order in W. (To make contact with the analysis of unitary coupled cluster theory in Refs. [27,28], write A_1 as $(T_1 - T_1^\dagger)$ and $A_2 = (T_2 - T_2^\dagger)$.) Consider now the expectation value of the energy $E = \langle e^{-A} H e^{A} \rangle$ *without* using any commutator approximations. Expanding in powers of the fluctuation operator we have

$E = E^0 + E^1 + E^2 + E^3 + E^4 + \cdots$, where the different order energies are defined as

$$E^0 = \langle E_{\mathrm{HF}} + F \rangle \tag{20a}$$

$$E^1 = \langle W \rangle \tag{20b}$$

$$E^2 = \langle [W, A_2] + [F, A_1] + [[F, A_2], A_2] \rangle \tag{20c}$$

$$E^3 = \left\langle \frac{1}{2}[[W, A_2], A_2] + [W, A_1] + \frac{1}{6}\Big[[[F, A_2], A_2], A_2\Big] + \frac{1}{2}[[F, A_2], A_1] + \frac{1}{2}[[F, A_1], A_2] \right\rangle \tag{20d}$$

$$\begin{aligned} E^4 = \bigg\langle & \frac{1}{6}\Big[[[W, A_2], A_2], A_2\Big] + \frac{1}{2}[[W, A_1], A_2] + \frac{1}{2}[[W, A_2], A_1] \\ & + \frac{1}{2}[[F, A_1], A_1] + \frac{1}{24}\Big[\Big[[[F, A_2], A_2], A_2\Big], A_2\Big] + \frac{1}{6}\Big[[[F, A_1], A_2], A_2\Big] \\ & + \frac{1}{6}\Big[[[F, A_2], A_1], A_2\Big] + \frac{1}{6}\Big[[[F, A_2], A_2], A_1\Big] \bigg\rangle. \end{aligned} \tag{20e}$$

Now consider the effect of the commutator approximation on the different orders of energy contribution. Firstly, no approximation is involved in computing E^0 or E^1, since these do not involve any commutators. For E^2, the approximation corresponds to

$$E^2 \Rightarrow \langle [W, A_2]_{1,2} + [F, A_1] + [[F, A_2], A_2]_{1,2} \rangle. \tag{21}$$

Note that no approximation is made for commutators like $[F, A_2]$ which generate only two- and lower-body operators. We see that $\langle [F, A_1] \rangle$ vanishes due to Brillouin's theorem, while both $[W, A_2]$ and $[[F, A_2], A_2]$ generate three-body operators that are neglected but have no expectation value with the Fermi vacuum and thus would not contribute to the energy. Thus no error is made in Equation (21) for E^2.

In the expression for E^3, we apply the commutator approximation twice for the double commutator $[[[W, A_2]_{1,2}, A_2]_{1,2}$. Once again, only fully contracted terms contribute to the energy. The only way fully contracted terms arise is from double contractions in $[W, A_2]$ to produce a two-particle operator, which then fully contracts with A_2 in the outer commutator and contributes to the energy. Since double contractions are involved in each step, the commutator approximation does not affect this term. There is no contribution from the three-body operators generated by either commutator, and so CT theory evaluates E^3 correctly also.

In the expression for E^4 we find our first error due to the commutator approximation. Here the three-body operator arising from the inner commutator $[W, A_2]$, which is neglected in the commutator approximation, can contract successively with two other A_2 terms in $[[[W, A_2], A_2], A_2]$ to yield a fully contracted term and thus a contribution to the energy. Although CT theory misses this contribution, it *does* contain the contribution that arises from contracting the two-particle operators generated in the inner commutator $[W, A_2]$. By a similar analysis, we find that the commutator approximation also provides an incomplete evaluation of $[[[[F, A_2], A_2], A_2], A_2]$, arising from intermediate three-body operators.

Could we avoid this error in E^4? It turns out that we can by using a more complicated *quadratic* commutator approximation $[[W, A_2], A_2] \rightarrow [[W, A_2], A_2]_{1,2}$, in which three- and four-body operators are approximated by one- and two-body operators after every *second* commutator. This approximation, while more complicated, does not change the computational scaling of the theory. Here the two offending terms are approximated as

$[[[W, A_2], A_2]_{1,2}, A_2]_{1,2}$ and $[[[[F, A_2], A_2], A_2]_{1,2}, A_2]_{1,2}$, where we neglect only the three- and four-body operators that result from the outer two commutators. As these operators do not contribute to any full contractions, they have no energy contribution. Thus by using a quadratic commutator approximation, CT theory's energy can be made correct through fourth order in the perturbation (although errors will appear at fifth order). This version of CT theory, termed QCTSD and discussed in detail in Section 5.3, is preferable when a single-determinant reference is employed.

In the usual coupled cluster hierarchy $\sum_{i=0}^{2} E_i$ is the MP2 energy functional, while $\sum_{i=0}^{3} E_i$ is the linearized coupled cluster single-doubles (L-CCSD) energy functional. $\sum_{i=0}^{4} E_i$ is the unitary CCSD energy functional. The CTSD energy is correct up to third order in perturbation theory (like linearized CCSD theory), while the QCTSD energy is correct up to fourth order (like CCSD). Unlike linearized CCSD theory, however, fourth order terms are not completely neglected in CTSD but are partly included as discussed above. From this, we might expect the single-reference CTSD theory to perform intermediate between linearized CCSD and the full CCSD theory. But in fact there are an *infinite* number of additional diagrams that are included in CTSD theory as compared to the usual CC and UCC(n) theories, because the energy functional does not terminate at finite order but contains further partial contributions from $E^5, E^6, \ldots, E^\infty$. For example, all terms involving pure orbital rotations (i.e. A_1) are included to *all* orders in the energy functional. Terms involving A_2, where all A_2 operators are at least doubly contracted with one other operator, are also included to all orders. One might speculate that these additional diagrams would yield an improved theory, but in the general case, and certainly when we extend the discussion to cases where a multi-determinantal reference wavefunction is used, the significance of the additional contributions present in CTSD (and QCTSD) can only be assessed numerically.

Finally we briefly analyse the case where the reference function is no longer a single Slater determinant. In this case errors arise not only from neglected operators, but also from neglected cumulants, the most significant being the three-particle cumulant. The nature of the error arising from neglecting the three-particle cumulant can be somewhat illuminated by assuming that the reference function itself admits a perturbation expansion in the active space. The corresponding fluctuation operator in the active space we will denote by W_{act}. Then

$$|\Psi_0\rangle = |\Phi^0\rangle + |\Phi^1\rangle + |\Phi^2\rangle + \cdots, \tag{22}$$

in which the zeroth order wavefunction $|\Phi^0\rangle$ is the Hartree–Fock solution. As the zeroth order wavefunction contains no correlation between electrons, it makes no contribution to the three-body cumulant. The structure of $|\Phi^1\rangle$ in Møller–Plesset theory is a sum of doubly excited determinants, which also gives no contribution to the three-body cumulant. In fact, it is not until we consider $|\Phi^2\rangle$, containing determinants with one-, two-, three-, and some four-body excitations, that we find a contribution to the three-body cumulant. We therefore see that if the Møller–Plesset expansion is a valid representation of our multi-configurational reference wavefunction, the three-body cumulant is of order W_{act}^2.

Let us now examine the contribution of the three-body cumulant to the energy. The three-body cumulant contribution arises in the following component of the first commutator of the BCH expansion, $\langle [W_{\mathrm{ext}}, A_2] \rangle$, where W_{ext} reflects the fluctuation potential between the active and external spaces. The operator A_2 is of order W_{ext}, so the product of amplitudes and integrals appearing in $\langle [W_{\text{ext}}, A_2] \rangle$ is of order W_{ext}^2, and its trace with the three-body cumulant yields an energy contribution of order $W_{\text{act}}^2 W_{\text{ext}}^2$. We see therefore that even in the multi-configurational case (at least in the case where the reference function itself admits a meaningful perturbative expansion), the error in the CTSD energy appears at fourth order in the fluctuation potentials.

3. Optimization

The excitation operator coefficients C_i that define the CT effective Hamiltonian are optimized by solving the CT amplitude equations, Equation (8). This set of nonlinear equations is analogous to the amplitude equations encountered in single-reference coupled cluster theory. While there are many methods available for solving sets of nonlinear equations, we employ the iterative Newton–Raphson (NR) method in CT theory for two reasons: first, the structure of the CT equations allows for a particularly efficient solution of the linear equation that defines the step in each NR iteration, and second, because the NR Jacobian matrix provides physical insight into why the CT equations are hard to solve and how they may be simplified. As mentioned in Section 2.1, the operators $\hat{o}_i$ that define our amplitude operator $A = \sum_i C_i \hat{o}_i$ are not simply taken to be the complete set of singles and doubles antisymmetric excitation operators given in Equation (3). For reasons explained below, the amplitude equations are unsolvable for this set of operators and so we employ more structured definitions of $\hat{o}_i$ instead.

3.1. *The Newton–Raphson approach*

When solving for the root of a single nonlinear equation $f(x)$, the NR approach is to make an initial guess $x^{(0)}$ and then extrapolate to the solution by following the function's slope $f'(x)$. This procedure leads to an iterative method in which the current solution is updated by the equation

$$x^{(i+1)} = x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})}. \tag{23}$$

This equation is evaluated repeatedly until the value of the function $f(x)$ is sufficiently small. The procedure for solving a system of nonlinear equations is analogous, except that now extrapolating to the solution using the functions' first derivatives requires solving a linear equation. The matrix in this equation, known as the Jacobian, consists of the derivative of each function with respect to each variable. In CT theory, our nonlinear equations are

$$R_i = \langle \Psi_0 | [\bar{H}, \hat{o}_i]_{1,2} | \Psi_0 \rangle = 0, \tag{24}$$

and so the Jacobian matrix is

$$J_{ij} = \frac{\partial R_i}{\partial C_j} = \langle \Psi_0 | [[\bar{H}, \hat{o}_j]_{1,2}, \hat{o}_i]_{1,2} | \Psi_0 \rangle + \mathrm{O}(A). \tag{25}$$

Given an initial guess $C_i^{(0)}$ for the excitation operator coefficients, one first evaluates the effective Hamiltonian $\bar{H}$ and determines the values R_i of the amplitude equations, which we often refer to as the residuals. An improved guess for the coefficients, $C_i^{(1)}$, is then found by solving the linear equation

$$\sum_j J_{ij} C_j^{(1)} = -R_i, \tag{26}$$

which extrapolates along the functions' derivatives in an analogous manner to Equation (23). Solving this linear equation is not trivial, however. Building the Jacobian is infeasible for all but the smallest systems, because it has a number of elements that scales as n^8 when the number of core, active, and virtual orbitals are assumed to be proportional to n. Instead we solve Equation (26) iteratively using a Krylov subspace method, which requires only the ability to compute the Jacobian matrix's action on a trial vector. This action can be evaluated efficiently if we neglect the last term of Equation (25), whose magnitude is linear in A and therefore should be small. Note that approximating the Jacobian in this way introduces no errors into the CT energy so long as the NR iteration leads to the solution of the amplitude equations. The NR equation can now be rewritten as

$$\left\langle \Psi_0 \middle| \left[\left[\bar{H}, \sum_j C_j^{(1)} \hat{o}_j \right]_{1,2}, \hat{o}_i \right]_{1,2} \middle| \Psi_0 \right\rangle = -R_i. \tag{27}$$

This form reveals how the Jacobian's action on the vector $C_j^{(1)}$ can be computed efficiently. The inner commutator of Equation (27) has the same form as the approximate commutator $[H, A]_{1,2}$ used to evaluate the BCH expansion and can therefore be evaluated in n^6 time. The result of this inner commutator is a two-body operator with the same symmetries as the electronic Hamiltonian, and so the outer commutator has precisely the same form as the amplitude equations, Equation (8), which can also be evaluated in n^6 time. Thus by evaluating and storing the intermediate operator resulting from the inner commutator of Equation (27), the action of the Jacobian on a trial vector can be evaluated efficiently and a Krylov subspace method can be used to solve Equation (26) in n^6 time.

A difficulty arises if any of the Jacobian's eigenvalues vanish or are very small, because linear equations involving singular or near singular matrices cannot in practice be solved using a Krylov subspace. Even if we could invert such a Jacobian, the result would not be physically meaningful, as one or more excitation operators would acquire a near-infinite coefficient. Recall that the working assumption for CT theory is that the reference function is close to the true wavefunction, which means that the excitation operators' coefficients should be small. As long as this assumption is satisfied (i.e. for a sufficiently large active space), any small eigenvalues in the Jacobian matrix are unphysical in nature and must be artifacts of CT theory's approximations. We call such eigenvalues and their corresponding excitation operators intruder states, because they are closely related to the intruder states encountered in second order perturbation theory (PT2) [55]. To see how, replace the effective Hamiltonian on the left hand side of Equation (27) with a zeroth order Hamiltonian satisfying $H_0|\Psi_0\rangle = E_0|\Psi_0\rangle$. By ignoring the commutator approximations, setting $C_i^{(0)} = 0$, and recalling that $\hat{o}_i^\dagger = -\hat{o}_i$, we may simplify the resulting equation to obtain

$$\sum_j \langle \Psi_0 | \hat{o}_i^\dagger (H_0 - E_0) \hat{o}_j | \Psi_0 \rangle C_j^{(1)} = \langle \Psi_0 | \hat{o}_i^\dagger H | \Psi_0 \rangle. \tag{28}$$

This is the defining equation for the coefficients $C_j^{(1)}$ of the perturber states $\hat{o}_j|\Psi_0\rangle$ that appear in the first order wavefunction of PT2 theory. Traditional intruder states in perturbation theory are those with unphysically large coefficients, and so we see that there is a strong analogy between intruder states in CT and PT2.

The key difference between intruder states in CT and PT2 theory is what causes them. For perturbation theory, Equation (28) tells us that intruder states will occur for inaccurate zeroth order Hamiltonians for which a state in the basis of perturber functions has an energy close to the reference state. In CT theory, Equation (25) contains no zeroth order Hamiltonian and so the origin of intruder states must be different. The three possibilities are errors in the effective Hamiltonian $\bar{H}$, neglecting the O(A) terms present in Equation (25), and the commutator approximations present in Equation (25). The first two possibilities can be ruled out if we consider the initial guess $C_i^{(0)} = 0$ (this is a reasonable guess for the coefficients as they should be small). For this guess the O(A) terms vanish and the initial effective Hamiltonian $\bar{H}$ is exactly equal to the bare electronic Hamiltonian H, which should give correct energies for states in the first order interacting basis. As CT theory typically uses $C_i = 0$ as its initial guess, any intruder states encountered are therefore caused by the commutator approximations present in Equation (25). Unfortunately, if one formulates CT theory using a multi-configurational reference function and the full set of excitation operators given in Equation (3), these commutator approximations create intruder states for all but the simplest systems. In the next two sections we discuss alternative choices for the excitation operators that are more successful at avoiding intruder states.

3.2. *Overlap truncation*

To address the problem of intruder states in CT theory, one must first recognize that for the standard choice of single and double excitation operators $\hat{p}_i \in \{a_q^p, a_{rs}^{pq} \ldots\}$, the first order interacting basis $\{\hat{p}_i|\Psi_0\rangle\}$ is not orthogonal if the reference is multi-configurational. Instead there is a dense, non-diagonal overlap matrix $S_{ij} = \langle\Psi_0|\hat{p}_i^\dagger \hat{p}_j|\Psi_0\rangle$. The eigenvalues of the Jacobian defined in Equation (25), which govern the presence of intruder states, are therefore represented by a generalized eigenvalue equation in which the overlap matrix appears:

$$\sum_j J_{ij} B_{jk} = \sum_l S_{il} B_{lk} \epsilon_k. \tag{29}$$

Here ϵ_k are the Jacobian eigenvalues and B is the matrix whose columns are the Jacobian eigenvectors. In order to improve the conditioning of the Newton–Raphson equation and remove linear degeneracies from the first order interacting basis, we can transform to the set of 'orthonormal' excitation operators $\hat{o}_i = \sum_j (S^{-1/2})_{ij} \hat{p}_j$ that generate an orthonormal first order interacting basis $\{\hat{o}_i|\Psi_0\rangle\}$. For these operators the Jacobian eigenvalues ϵ_k are defined by the simple eigenvalue equation

$$\sum_j \tilde{J}_{ij} \tilde{B}_{jk} = \tilde{B}_{ik} \epsilon_k, \tag{30}$$

in which $\tilde{J} = S^{-1/2} J S^{-1/2}$ and $\tilde{B} = S^{1/2} B$ are the Jacobian and its eigenvectors in the orthonormal basis $\{\hat{o}_i|\Psi_0\rangle\}$. To prevent intruder states, the small eigenvalues present in S and J must cancel in the product $S^{-1/2}JS^{-1/2}$. Errors in the eigenvalue spectrums of either S or J can disrupt this cancellation and produce unphysically small values for ϵ_k, which as described in the previous section will correspond to intruder states. There are two possible sources for such errors. First, for semi-internal excitation operators (double excitations that involve one excitation into or out of the active space and one excitation within the active space) the overlap matrix S depends on the reference function's three-body RDM. If we approximate S by neglecting the three-body cumulant in order to avoid using the three-body RDM, then we introduce errors in S that can lead to intruder states. Second, the commutator approximations present in the definition of J create errors in its eigenvalue spectrum. In practice we find that even if the exact three-body RDM is used to avoid errors in S, the errors in J are sufficient to produce intruder states for most multi-configurational references. Therefore, in order to circumvent the need for a delicate cancellation of small eigenvalues in the product $S^{-1/2}JS^{-1/2}$, we restrict our operators $\hat{o}_i$ by discarding the eigenvalues of S below some threshold τ. In practice we use two thresholds, τ_1 for the single and semi-internal excitations' overlap matrices and τ_2 for the double excitations' overlap matrices. While all multi-configurational dynamic correlation methods employing an internally contracted form must truncate the overlap matrices to some extent (e.g. $\tau = 10^{-6}$) in order to remove true linear degeneracies from the first order interacting basis, CT theory requires a much larger truncation threshold (e.g. $\tau = 10^{-2}$) in order to prevent intruder states. Because of the importance of accurately truncating the overlap matrices, the standard CTSD version of CT theory, reported previously as L-CTSD(MK) [2] and LCTSD [4], uses the exact three-body RDM in order to compute the semi-internal overlap matrices. We note that CTSD does not, however, use the three-body RDM anywhere else.

The use of overlap truncation in CT theory has been very successful, producing accuracies competitive with expensive methods such as MRCI+Q. However, the need to accurately truncate the overlap matrices creates three significant disadvantages. First, the choice of the truncation thresholds is arbitrary and in difficult systems can affect the CT energy, as discussed in Section 6.3. Second, constructing the overlap matrices exactly with the reference wavefunction's three-body RDM is necessary to achieve these accuracies. While overlap truncation can use the cumulant-approximated overlap matrices, the accuracy suffers because higher truncation thresholds are necessary to prevent the errors in S from creating intruder states. Finally, diagonalizing the semi-internal overlap matrices to produce $S^{-1/2}$ has a cost that scales as n_{act}^9, where n_{act} is the number of active orbitals. This cost is trivial for small active spaces, but will become infeasible for the very large active spaces accessible with DMRG theory. There is another way to construct the operator set $\{\hat{o}_i\}$ that avoids these issues at the cost of further limiting the ansatz's freedom, which we shall now discuss.

3.3. *Strong contraction*

Strongly contracted (SC) excitation operators were first introduced by Malrieu *et al.* in the context of n-electron valence perturbation theory (NEVPT2) [16,17]. They consist of a drastic simplification of the first order interaction basis in which each external orbital (for singles) or orbital pair (for doubles) has only one excitation operator that connects it to the active space. Such an operator is of course a linear combination of many of the basic excitation operators, but all terms in the combination involve the same external indices. An immediate consequence of this formulation is that the SC operators are mutually orthogonal. These operators thus avoid completely the difficulties of building, diagonalizing, and truncating overlap matrices and therefore require neither the n_{act}^9 cost diagonalization step nor the reference function's three-body RDM. While SC operators are certainly simple to work with, one must construct them carefully in order to retain accuracy in such a restricted set of excitations. Each SC operator is therefore formed as the sum of all contracted operators of its type (e.g. double excitations from the active space into the virtual orbitals v_1 and v_2) weighted by their coefficients in the electronic Hamiltonian H. For the example of double excitations between the active and virtual orbitals, the operator corresponding to the pair of virtual orbitals (v_1, v_2) is

$$\hat{o}^{v_1 v_2} = \sum_{a_1 a_2} g_{a_1 a_2}^{v_1 v_2} \left(a_{a_1 a_2}^{v_1 v_2} - a_{v_1 v_2}^{a_1 a_2} \right), \tag{31}$$

where g is the usual two-body integral tensor. In total there are eight types of SC operators, the precise definitions for which are given in the appendix of Ref. [5]. The justification for weighting the operators based on their Hamiltonian coefficients is that it makes the states $\hat{o}_i|\Psi_0\rangle$ components of the system's first Krylov vector, $H|\Psi_0\rangle$. The operation of H on the reference function emphasizes the eigenstates of H in $|\Psi_0\rangle$ with the largest magnitude eigenvalues. As states with large positive eigenvalues will not be present in $|\Psi_0\rangle$, the emphasis will be put on those eigenstates with large negative eigenvalues, namely the ground state and low-lying excited states. Thus the choice of Hamiltonian coefficients for constructing a SC excitation operator is good in the sense that it is tailored towards describing the system's lowest eigenstates. Even with this intelligent choice for the form of the SC operators, they offer far less freedom than is available through the overlap truncation method, and the accuracy of CT suffers as a result. The motivation for using SC operators is of course not to improve accuracy (which for CTSD was already excellent) but to avoid intruder states.

In practice, the orthogonality of the SC operators is sufficient to prevent intruder states in many systems. To detect and remove any remaining intruder states, we construct an approximation of the CT Jacobian and inspect it for any unphysically small eigenvalues. This approximation consists of neglecting the last term in Equation (25) and replacing the effective Hamiltonian with Dyall's zeroth order Hamiltonian [56],

$$H_0 = C + \sum_{c_1} \bar{t}_{c_1}^{c_1} a_{c_1}^{c_1} + \sum_{v_1} \bar{t}_{v_1}^{v_1} a_{v_1}^{v_1} + \sum_{a_1 a_2} \bar{t}_{a_2}^{a_1} a_{a_2}^{a_1} + \frac{1}{2} \sum_{a_1 a_2 a_3 a_4} g_{a_3 a_4}^{a_1 a_2} a_{a_3 a_4}^{a_1 a_2}, \tag{32}$$

in which C is a constant, $\bar{t}$ is a set of effective one-body integrals, and g is the usual two-body integral tensor. Note that C and $\bar{t}$ are defined such that $H_0|\Psi_0\rangle = E_0|\Psi_0\rangle$. Our approximation for the CT Jacobian is now

$$J_{ij} = \langle \Psi_0 | [[H_0, \hat{o}_j]_{1,2}, \hat{o}_i]_{1,2} | \Psi_0 \rangle, \tag{33}$$

which is a diagonal matrix because the Dyall Hamiltonian cannot connect operators that excite into different external orbitals (recall that each SC operator corresponds to a different set of external orbitals). As in the previous section, we desire the Jacobian eigenvalues in an orthonormal basis of excitation operators, and although the SC operators are orthogonal by construction they are not normalized. We therefore evaluate their norms approximately via

$$||\hat{o}_i||^2 \simeq \langle\Psi_0|\hat{o}_i^\dagger \hat{o}_i|\Psi_0\rangle_{1,2}, \tag{34}$$

where we neglect the three-body cumulant to avoid using the three-body RDM. With these approximate norms, we may evaluate the approximate eigenvalues of the orthonormal Jacobian as

$$\epsilon_i = \frac{1}{||\hat{o}_i||^2}\langle\Psi_0|[[H_0, \hat{o}_i]_{1,2}, \hat{o}_i]_{1,2}|\Psi_0\rangle. \tag{35}$$

After obtaining these approximate Jacobian eigenvalues (which can be done in n^6 time), we inspect them and remove from our excitation operator basis any operators with unphysically small values for ϵ_i. These eigenvalues should be bounded from below by the Dyall Hamiltonian's smallest excitation energy between the reference state and states in the first order interacting basis. For a sufficiently large active space, these excitation energies will be quite sizable as the cost to occupy a virtual orbital or create a hole among the core orbitals will be high. In practice, we typically discard operators with ϵ_i below a threshold of $\tau_\epsilon = 0.1$ Hartrees. This truncation threshold is more physically intuitive than the overlap truncation thresholds presented in the previous section, and in large active spaces the CT energy is insensitive to its value (e.g. all excitation operators have ϵ_i much larger than 0.1 Hartrees). In summary, SC operators prevent intruder states in CT theory without building or diagonalizing overlap matrices or using the three-body RDM.

4. The parallel CT algorithm

The most computationally demanding procedures in a CT calculation are to evaluate the approximate commutator, $[H, A]_{1,2}$, and the values of the amplitude equations, $R_i = \langle\Psi_0|[H, \hat{o}_i]_{1,2}|\Psi_0\rangle$. We approach this task using a parallel algorithm which distributes the floating point operations and some of the storage across multiple computer processors. Before describing the exact structure of this algorithm, let us specify clearly the operators and tensors involved in the implementation. For efficiency, we work in the spin free form of second quantization, which is based on the so-called unitary group *generators* given by

$$\hat{E}^{p_1}_{q_1} = \sum_{\sigma=\alpha,\beta} a^\dagger_{p_1\sigma} a_{q_1\sigma}, \tag{36a}$$

$$\hat{E}^{p_1p_2}_{q_1q_2} = \sum_{\sigma\tau=\alpha,\beta} a^\dagger_{p_1\sigma} a^\dagger_{p_2\tau} a_{q_2\tau} a_{q_1\sigma}, \tag{36b}$$

$$\hat{E}^{p_1p_2p_3}_{q_1q_2q_3} = \sum_{\sigma\tau\upsilon=\alpha,\beta} a^\dagger_{p_1\sigma} a^\dagger_{p_2\tau} a^\dagger_{p_3\upsilon} a_{q_3\upsilon} a_{q_2\tau} a_{q_1\sigma}, \tag{36c}$$

for the one-, two-, and three-body operators, respectively. The reduced density matrices (RDMs) are accordingly given in the spin free form as the expectation values of these generators,

$$D^{p_1}_{q_1} = \langle\Psi_0|\hat{E}^{p_1}_{q_1}|\Psi_0\rangle, \qquad D^{p_1p_2}_{q_1q_2} = \langle\Psi_0|\hat{E}^{p_1p_2}_{q_1q_2}|\Psi_0\rangle, \qquad D^{p_1p_2p_3}_{q_1q_2q_3} = \langle\Psi_0|\hat{E}^{p_1p_2p_3}_{q_1q_2q_3}|\Psi_0\rangle. \tag{37}$$

In this spin free formulation, the CT amplitude operator may be written as a sum of a one-body and a two-body operator,

$$A = A_1 + A_2, \tag{38a}$$

$$A_1 = \sum_{e_1 a_1} A^{e_1}_{a_1}(\hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}), \tag{38b}$$

$$A_2 = \frac{1}{2}\sum_{e_1 e_2 a_1 a_2} A^{e_1 e_2}_{a_1 a_2}(\hat{E}^{e_1 e_2}_{a_1 a_2} - \hat{E}^{a_1 a_2}_{e_1 e_2}), \tag{38c}$$

regardless of the form chosen for the operators $\hat{o}_i$. Finally, a general Hamiltonian (either the electronic Hamiltonian or a canonically transformed effective Hamiltonian), can be written in the form

$$H = h_0 + h_1 + h_2, \tag{39a}$$

$$h_1 = \sum_{p_1 p_2} t^{p_1}_{p_2}\hat{E}^{p_1}_{p_2}, \tag{39b}$$

$$h_2 = \frac{1}{2}\sum_{p_1 p_2 p_3 p_4} v^{p_1 p_2}_{p_3 p_4}\hat{E}^{p_1 p_2}_{p_3 p_4}, \tag{39c}$$

where h_0 is a constant and t and v are the one- and two-body integral tensors, respectively. With these definitions in hand, we may now proceed to describe our algorithm.

Our central parallelization strategy is to distribute the storage of the two-body integral tensor $v^{p_1 p_2}_{p_3 p_4}$ across our processors. This strategy is motivated by the n^4 size (where n is the number of total orbitals) of this tensor, which rapidly becomes too large to store in the fast memory of a single processor. We do so by splitting the tensor by its p_1 index, so that each processor stores the integrals for a limited set of values of p_1 and all values of p_2, p_3, and p_4. To further reduce memory requirements, we store these integrals on the hard disk of each processor, loading them into fast memory one n^3 sized block at a time. A block is defined as the set of all two-body integrals corresponding to a specific value of p_1. As we shall see, this distribution and blocking of the two-body integrals allows us to evenly distribute CT theory's floating point operations across our processors. The remaining tensors (h_0, $h^{p_1}_{p_2}$, $D^{p_1}_{p_2}$, $D^{p_1 p_2}_{p_3 p_4}$, $A^{e_1}_{a_1}$, $A^{e_1 e_2}_{a_1 a_2}$, $R^{e_1}_{a_1}$, $R^{e_1 e_2}_{a_1 a_2}$) are stored redundantly in the fast memory of each processor. The worst of these tensors, $A^{e_1 e_2}_{a_1 a_2}$ and $R^{e_1 e_2}_{a_1 a_2}$, are of size $n^2_{\text{occ}} n^2_{\text{open}}$, where n_{occ} is the number of core and active orbitals while n_{open} is the number of active and virtual orbitals.

If we temporarily set aside the terms $[h_2, A_1]$ and $[h_2, A_2]_{1,2}$, all other terms in the approximate commutator and amplitude equations ($[h_1, A_1]$, $[h_1, A_2]$, $\langle\Psi_0|[h_1, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}]|\Psi_0\rangle$, $\langle\Psi_0|[h_2, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}]|\Psi_0\rangle$, $\langle\Psi_0|[h_1, \hat{E}^{e_1 e_2}_{a_1 a_2} - \hat{E}^{e_1 e_2}_{a_1 a_2}]|\Psi_0\rangle$, and $\langle\Psi_0|[h_2, \hat{E}^{e_1 e_2}_{a_1 a_2} - \hat{E}^{e_1 e_2}_{a_1 a_2}]_{1,2}|\Psi_0\rangle$) can now be evaluated in parallel. We see that all tensors in these terms except the two-body integrals are stored redundantly on each processor. The necessary tensor contractions (given in the appendix) can therefore be carried out for each block of two-body integrals independently, and so each processor does not need any information from the others to evaluate its contribution. Once all processors have evaluated their individual contributions, these are summed together. This final summation requires an $n^2_{\text{occ}} n^2_{\text{open}} \log n_{\text{proc}}$ amount of data transfer between processors, where n_{proc} is the number of processors.

Evaluating the terms $[h_2, A_1]$ and $[h_2, A_2]_{1,2}$ is more challenging, as they each have two sets of two-body integrals. For now we will label these as $v^{p_1p_2}_{p_3p_4}$ for the tensor inside h_2 and $w^{q_1q_2}_{q_3q_4}$ for the tensor inside the two-body operator resulting from the commutator. At first glance, evaluating these terms requires access to all blocks of w for each block of v. However, a careful inspection of these terms and the four-fold symmetries of v and w ($v^{p_1p_2}_{p_3p_4} = v^{p_2p_1}_{p_4p_3} = v^{p_3p_4}_{p_1p_2} = v^{p_4p_3}_{p_2p_1}$) reveals that each block of v can be made to contribute only to the same block of w, as long as w is symmetrized afterwards. This means that each processor can evaluate these terms for its set of integral blocks without needing any information from the other processors. We may therefore evaluate these terms through the following process, for which the necessary tensor contractions are given in the appendix.

(1) The orbital range for p_1 is divided evenly into sub-ranges, with each sub-range assigned to a different processor. Thus the two-body integrals $v^{x_1p_2}_{p_3p_4}$ $\forall$ p_2, p_3, p_4 are stored on the hard disk of the processor assigned x_1.
(2) On each processor, select x_1 from the assigned range.
(3) Load the n^3 sized block of two-body integrals $v^{x_1p_2}_{p_3p_4}$ $\forall$ p_2, p_3, p_4 into the processor's fast memory.
(4) Evaluate the tensor contractions to produce the block of integrals $w^{x_1q_2}_{q_3q_4}$ $\forall$ q_2, q_3, q_4.
(5) Write the integral block $w^{x_1q_2}_{q_3q_4}$ $\forall$ q_2, q_3, q_4 to the processor's hard disk.
(6) Go to step 2 until all values of x_1 belonging to this processor have been exhausted.
(7) Symmetrize the integrals w across the processors. This step requires an n^4/n_{proc} amount of data transfer.

In summary, the floating point operations required for all terms in $[H, A]_{1,2}$ and $\langle\Psi_0|[H, \hat{o}_i]_{1,2}|\Psi_0\rangle$ are evenly distributed across all processors. This results in a fully parallel CT algorithm whose per-processor computational requirements amount to $O(n^2_{\text{occ}}n^4_{\text{open}}/n_{\text{proc}})$ floating point operations, $O(n^3) + O(n^2_{\text{occ}}n^2_{\text{open}})$ storage in fast memory, $O(n^4/n_{\text{proc}})$ data transfer to and from hard disk, and $O(n^2_{\text{occ}}n^2_{\text{open}} \log n_{\text{proc}}) + O(n^4/n_{\text{proc}})$ data transfer to and from other processors.

5. Special topics

5.1. *CT and DMRG*

One of the primary motivations driving the development of CT theory, as discussed in the introduction, is the need for a dynamic correlation method that can be used in conjunction with the description of static correlation in large active spaces. We have not yet, however, specified how the static correlation in such large active spaces is to be obtained: the commonly used CASSCF method for treating static correlation cannot treat systems with more than 16 active orbitals. A natural candidate for static correlation in large active spaces is the DMRG, which in work by our group and others [57–80] has been shown to accurately describe static correlation in general molecules with up to 40 active orbitals, and linear molecules with as many as 100 active orbitals. Evaluating the three-body RDM in DMRG theory is more expensive than obtaining the DMRG wavefunction itself, and so in most systems the DMRG three-body RDM is unavailable. CT theory (especially the strongly contracted variant) can make do with only the reference wavefunction's one- and two-body RDMs and is therefore ideally suited to work with a DMRG wavefunction.

While the combination of these two theories has been envisioned for some time, only recently has their use together become practical. So far, they have been used to evaluate the singlet-triplet gap in free base porphin (24 active orbitals) [5], predict accurate correlation energies for long polyenes (24 active orbitals) [6], and model the isomerization of the $[Cu_2O_2]^{2+}$ complex (32 active orbitals) [6]. These applications are reviewed in Section 6.5

5.2. *Excited states*

Although the use of CT theory to evaluate excitation energies is not fully developed, we include a short discussion of how it can be done for those readers who are interested. In general, a multi-configurational dynamic correlation method may attack the problem of excited states in one of two ways. The first option is to produce two separately optimized reference states using the underlying static correlation method (CASSCF, DMRG-SCF, etc.), one for the ground state and the other for the excited state in question. Two completely separate CT effective Hamiltonians are then found by solving the separate CT amplitude equations, using the ground state's RDMs in one calculation and the excited state's in the other. The final energies are then subtracted to obtain the energy of excitation. The other option for calculating excited states in CT theory is to use some form of state averaging, and it is this option which we have recently explored [81]. For simplicity, assume that our reference functions are a set of CASSCF solutions in a common set of molecular orbitals, which have been optimized using the average density matrix of the CASSCF states in question. We then attempt to satisfy as best we can the CT amplitude equations for all states simultaneously by solving them using RDMs that are the average of the individual states' RDMs. This produces an amplitude operator A that defines one canonical transformation e^A for all states. Using this state-averaged amplitude operator A, we are then faced with two choices for how to define the effective Hamiltonian appropriate to the excited states, due to the use of operator decompositions (which themselves rely on state information through the density matrices) in the BCH expansion. The first choice is to construct a single effective Hamiltonian common to all states using a state-averaged RDM in all the operator decompositions. The second is to create an effective Hamiltonian for each state specifically by evaluating our operator decompositions with the state-specific RDMs. Either way, the excited state energies are evaluated by tracing the effective Hamiltonian(s) with each state's individual density matrices.

5.3. *Quadratic CT theory*

As discussed in Section 2.5, standard CT theory with singles and doubles (CTSD) is formally accurate to lower order in perturbation theory in single-reference systems than CCSD theory. This disadvantage may be removed in the following way. In the standard theory, the 'quadratic' commutator $[[H, A], A]$ is replaced by a term where the operator decomposition is applied twice, $[[H, A]_{1,2}, A]_{1,2}$, which we have previously referred to as a linear commutator approximation [4]. If instead, we delay the operator decomposition until after the second commutator, we obtain a 'quadratic' commutator approximation,

$[[H,A],A]_{1,2}$, which when used in the BCH expansion yields a quadratic CT theory (QCTSD) that is accurate to the same order as CCSD [4].

The quadratic commutator approximation requires a decomposition for four-body operators in terms of one- and two-body operators in addition to the decomposition for three-body operators discussed in Section 2.2 and employed by CTSD. Like its three-body counterpart, the four-body decomposition is derived from extended normal ordering (ENO) by neglecting all cumulants and ENO operators of particle ranks 3 and higher. The explicit form of the decomposition is given in Ref. [4].

In the QCTSD method, the CT equations are rewritten to delay operator decomposition for as long as possible. In order to do so, it becomes necessary to treat the one- and two-body portions (h_1 and h_2) of the Hamiltonian separately when evaluating the BCH expansion so that operator decomposition for h_1 can be delayed to one order higher in A. The resulting form of $\bar{H}$ used in QCTSD is therefore

$$\bar{H} = H + \bar{H}^{(1)} + \bar{H}^{(2)}, \tag{40a}$$

$$\bar{H}^{(1)} = [h_1, A] + \frac{1}{2!}[[h_1, A], A]_{1,2} + \frac{1}{3!}[[[h_1, A], A], A]_{1,2} + \frac{1}{4!}[[[[h_1, A], A], A]_{1,2}, A]_{1,2} + \cdots, \tag{40b}$$

$$\bar{H}^{(2)} = [h_2, A]_{1,2} + \frac{1}{2!}[[h_2, A], A]_{1,2} + \frac{1}{3!}[[[h_2, A], A]_{1,2}, A]_{1,2} + \frac{1}{4!}[[[[h_2, A], A]_{1,2}, A], A]_{1,2} + \cdots \tag{40c}$$

Similarly, the amplitude equations must distinguish between the one- and two-body components of the effective Hamiltonian ($\bar{h}_1$ and $\bar{h}_2$) in order to delay operator decompositions for as long as possible.

$$R_i = R_i^{(1)} + R_i^{(2)} \tag{41a}$$

$$R_i^{(1)} = \langle[\bar{h}_1, \hat{o}_i]\rangle + \langle[[\bar{h}_1, A], \hat{o}_i]_{1,2}\rangle + \frac{1}{2!}\langle[[[\bar{h}_1, A], A], \hat{o}_i]_{1,2}\rangle + \frac{1}{3!}\langle[[[[\bar{h}_1, A], A], A]_{1,2}, \hat{o}_i]_{1,2}\rangle + \cdots \tag{41b}$$

$$R_i^{(2)} = \langle[\bar{h}_2, \hat{o}]_{1,2}\rangle + \langle[[\bar{h}_2, A], \hat{o}]_{1,2}\rangle + \frac{1}{2!}\langle[[[\bar{h}_2, A], A]_{1,2}, \hat{o}]_{1,2}\rangle + \frac{1}{3!}\langle[[[[\bar{h}_2, A], A]_{1,2}, A], \hat{o}_i]_{1,2}\rangle + \cdots \tag{41c}$$

A final difference between QCTSD and CTSD is the evaluation of the Jacobian's action, for which QCTSD uses the quadratic commutator approximation rather than two successive linear commutator approximations,

$$\sum_j J_{ij} C_j = \langle\Psi_0|[[\bar{H}, \sum_j C_j \hat{o}_j], \hat{o}_i]_{1,2}|\Psi_0\rangle. \tag{42}$$

While QCTSD improves the formal accuracy of CTSD for single-determinant reference functions, it is in practice less accurate than CTSD when the reference function consists of many determinants. The reasons behind this loss of accuracy are not clear and should perhaps be analyzed more carefully in future studies. As a final note, QCTSD is more expensive than CTSD due to the much larger number of tensor contractions required to evaluate the quadratic commutator approximation. QCTSD does, however, retain the desirable n^6 cost scaling and is therefore more expensive than CTSD by a constant factor only.

5.4. *The three-body RDM*

While one of CT theory's most important properties is the ability to treat reference functions for which the three-body RDM is not available, many important chemical systems have small enough active spaces that producing this RDM is affordable. It is desirable that for such systems CT theory be able to make use of this additional wavefunction information in order to improve its accuracy. Indeed, it has been shown [4] that the accuracy of results for the bond breaking of N_2 and H_2O can be improved by simply redefining the commutator approximation $[H, A]_{1,2}$ so that the three-body cumulant is not discarded. Furthermore, access to the exact three-body RDM significantly improves the effectiveness of the overlap truncation method at removing intruder states, which provides an even larger accuracy improvement than the improved commutator approximation. One place where the effect of using the exact three-body RDM rather than its cumulant approximation has not been tested is in normalizing and constructing the Dyall-approximated Jacobian for strongly contracted excitation operators. We expect that this will create similar improvements to those seen for overlap truncation, and this would be interesting to explore in the future. The chief drawback of using the three-body RDM is that it increases the cost of evaluating the commutator approximation to n^7. However, systems for which the three-body RDM can be obtained are likely to have small enough active spaces so that this additional cost is affordable. In summary, for systems in which the three-body RDM is available, its use should be incorporated into CT theory as much as possible in order to improve accuracy.

5.5. *Automatic derivation*

The most difficult part of implementing a CT algorithm is deriving and encoding the tensor contractions necessary to compute the commutator approximations. The scale of this task becomes clear when one considers that the approximate quadratic commutator $[[H, A], A]_{1,2}$ that appears in QCTSD theory contains over 16,000 unique terms [4]. For many of the expressions present in CT theory, we therefore use a computer program to automate the derivation process. This program may be downloaded from the erratum of Ref. [4]. The central tasks necessary to derive an efficient CT algorithm are: expanding all commutators, normal ordering creation and destruction operators, replacing many-body operators by their operator decompositions, combining like terms, and finally choosing the most efficient order in which to contract the tensors in each term. The program in question performs all but the last of these tasks, which we accomplish through a separate, unpublished program whose operation we will describe below.

The most difficult step in deriving the terms involved in a commutator approximation is comparing terms to determine if they may be combined. Term comparison is made difficult by the symmetry of the tensors and the freedom to rename dummy indices.

As an example, consider the term

$$\sum_{\substack{a_1a_2a_3\\ i_1i_2i_3i_4i_5}} \gamma^{a_3}_{i_3}\gamma^{i_4a_1}_{i_1i_5}v^{i_2i_4}_{i_3i_5}A^{a_1}_{a_2}A^{a_2a_3}_{i_2i_1} \tag{43}$$

which occurs when deriving the explicit expression for $[[H, A], A]_{1,2}$. The tensors in this term have two-, eight-, eight-, one-, and four-fold symmetries for a total of 512 possible index arrangements. Further, because of the summation, all the indices are dummy indices. In order to combine like terms, the program first transforms each term into a unique canonical form (here the word canonical has no relation to CT theory). The rules for writing a term in canonical form are based on a lexicographic ordering of the tensors and their indices, with some special rules in the case of a repeated tensor name. This choice of canonical form is arbitrary and certainly not unique. What is important is that for each term encountered in CT theory, there is only one way to write it in canonical form. The canonical form for our example term is

$$\sum_{\substack{abcd\\ efgh}} A^a_bA^{bc}_{de}\gamma^c_f\gamma^{ag}_{dh}v^{eg}_{fh}. \tag{44}$$

The comparison of terms may be seen as a problem of graph isomorphism in which the tensors represent labeled vertices and the summation indices labeled edges. As the general problem of graph isomorphism has no known solutions of polynomial cost, we expect that our problem of term combination will become unmanageable for a sufficiently large number of tensors. However, the terms involved in CT theory are simple enough that the cost of converting them to canonical form is not prohibitive. We do note, though, that it is the most expensive step in the automatic derivation process.

After all of the necessary terms have been derived, it is important to choose an efficient order in which to carry out the contractions for a specific term. In our example term, one must be careful not perform the n^7 cost contraction $\sum_e A^{bc}_{de}v^{eg}_{fh}$. Instead, by first contracting A^a_b and γ^c_f with A^{bc}_{de}, the contractions with γ^{ag}_{dh} and v^{eg}_{fh} can be completed for only an n^6 cost. While one could exhaustively search all possible contraction orders to guarantee maximal efficiency, this process has a potentially unaffordable $N!$ cost for N tensors. Instead, our program looks for the least expensive contraction pair out of all the tensors and chooses this as the first contraction, then repeats this process until all tensors have been assigned a position in the contraction ordering. While this process is not guaranteed to find the most efficient contraction ordering, it can be completed quickly as it has only an N^3 cost. In practice, we tell the program to warn us if it finds a term for which this ordering would produce a contraction cost higher than n^6, which is our target cost for CT theory. Such terms occur infrequently and are inspected manually in order to find a more efficient contraction ordering. After all terms are derived and all contraction orderings chosen, the necessary cache-optimized FORTRAN code is written automatically.

6. Results and applications

6.1. *A small molecule example: N_2*

Correctly breaking the bond of the nitrogen dimer is a particularly challenging problem for quantum chemical theories due to the strong role of static correlation. Near equilibrium, the N_2 wavefunction is dominated by a single configuration in which the bonding π orbitals are doubly occupied. However, as the bond is broken, excited configurations play an increasingly important role, leading to the qualitative failure of single-reference methods. The use of a multireference method such as CASSCF, which includes all possible configurations of the electrons in the 2p orbitals, restores qualitative agreement with the exact theoretical (FCI) result, but still fails to achieve quantitative agreement due to the omission of dynamic correlation. By quantitative agreement we refer to chemical accuracy, commonly regarded as a relative error of less than 1 kcal/mol. In a binding curve the relative error refers to the worst case error in energy differences between points on the curve, so we may think of it as measuring the parallelity of the computed curve to the exact curve. Significantly, not all multireference dynamic correlation theories are able to achieve chemical accuracy in the N_2 binding curve, despite its small size and the affordability of constructing and diagonalizing its three-body RDM. Most notable is the failure of perturbation theory, which for example produces a 5.2 kcal/mol relative error when implemented as CASPT2, as shown in Table 1 and Figure 1. In contrast, the two standard CT methods, CTSD with exact overlap and SC-CTSD, have relative errors of 0.6 and 1.0 kcal/mol, respectively. Both of these methods have lower cost scalings than CASPT2 and can therefore be applied to much larger systems. Of particular note is that the SC-CTSD method achieves chemical accuracy in N_2 using only the one- and two-body RDMs of the reference wavefunction, an unprecedented feat. Other methods with comparable accuracy, such as MRCI+Q and CASPT3, require access to either the CASSCF wavefunction's CI coefficients or the three-body RDM.

The breaking of the N_2 bond has been extensively studied with CT theory and provides an excellent pedagogical example of the theory's strengths and weaknesses. The relatively poor performance of CTSD when using the cumulant-approximated overlap matrices is a reminder of CT theory's intruder states. The accuracy is damaged in this case by the necessity of using larger truncation thresholds to prevent errors in the overlap matrices

Table 1. Results for the N_2 potential energy curve. The FCI total energy is reported in E_h, with other methods reported as the difference from FCI in mE_h. See Section 6.1 and Figure 1 for details.

R[a]	FCI[b]	CASSCF	MRCI+Q	CASPT2	CTSD[c]	CTSD[d]	SC-CTSD[e]
0.9525	−109.167573	182.072	−0.564	21.437	1.883	−0.073	5.850
1.0679	−109.270384	186.030	−0.782	22.202	2.423	−0.008	5.668
1.1208	−109.278339	187.644	−0.845	22.497	2.448	0.352	5.723
1.1737	−109.271915	189.155	−0.912	22.713	1.801	−0.519	5.713
1.2700	−109.238397	191.715	−1.029	22.873	2.940	−0.535	5.497
1.4288	−109.160305	195.376	−1.174	22.300	5.776	−0.015	5.428
1.5875	−109.086211	197.685	−1.363	20.309	5.019	−0.272	4.344
1.7463	−109.03031	197.660	−1.491	17.287	1.709	0.053	5.837
1.9050	−108.99481	194.696	−1.750	14.587	0.339	−0.482	4.202
Relative error (mE_h)		15.612	1.186	8.286	5.437	0.887	1.648

[a] Radius in angstroms.

[b] Taken from Ref. [90].

[c] $\tau_1 = 0.3$, $\tau_2 = 0.1$, approximate overlap.

[d] $\tau_1 = 0.1$, $\tau_2 = 0.01$, exact overlap.

[e] $\tau_\epsilon = 0.1\ E_h$.

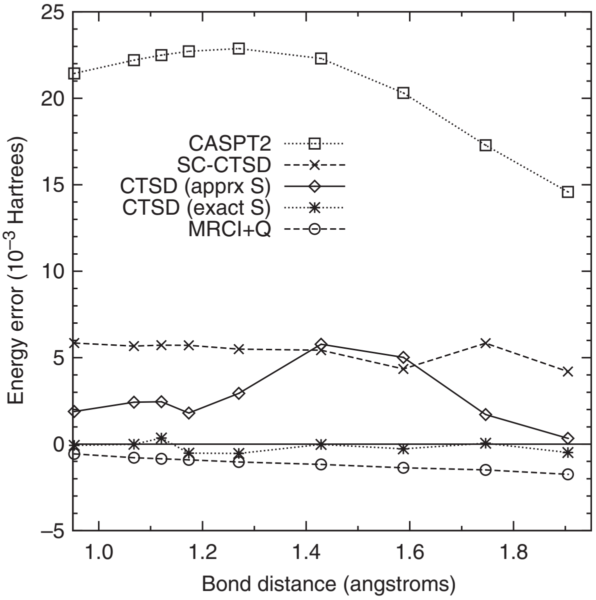

Figure 1. Ground state energy errors relative to FCI for N_2 in the cc-pVDZ basis set using spherical d orbitals. All methods use the six 2p orbitals as their active space, with all orbitals optimized during the CASSCF calculation. The 1s orbitals are held frozen during the dynamic correlation calculations. CTSD employed truncation thresholds of ($\tau_1 = 0.3$, $\tau_2 = 0.1$) and ($\tau_1 = 0.1$, $\tau_2 = 0.01$) when using the approximate and exact semi-internal overlap matrices, respectively. For SC-CTSD, an energy threshold of $\tau_\epsilon = 0.1$ E_h was used. In addition, the strongly contracted $2s\sigma^* \rightarrow 3s\sigma^*$ semi-internal excitation displayed intruder state character and was disabled manually.

from creating intruder states. The result is a relative error of 3.4 kcal/mol, more than five times larger than the 0.6 kcal/mol error produced when the exact three-body RDM is used to construct the overlap matrices. Strongly contracted operators restrict the ansatz's freedom in a more intelligent way than the approximate overlap truncation and are effective at removing intruder states without using the three-body RDM. SC-CTSD's relative error of 1.0 kcal/mol is a significant improvement versus 3.4 kcal/mol and is almost as good as the 0.6 kcal/mol achieved by using the exact three-body RDM. This hierarchy of accuracy (CTSD with three-body RDM > SC-CTSD > CTSD without three-body RDM) is typical and has also been observed in H_2O and NiO.

6.2. *Size consistency*

As discussed in Section 2.4, CT theory is rigorously size consistent. In this section we shall summarize previous results comparing the true size consistency of CT theory with the approximate size consistency of some other methods, as well as present new data to enhance the comparison. To measure its accuracy in systems too large to be treated using FCI, CT theory is often compared against the ACPF, AQCC, and MRCI+Q methods, which are configuration interaction methods that have been modified to approximately

Table 2. Size consistency errors (absolute values) for various systems and methods in mE_h. Entries denoted with – were not calculated. Note that for a Hartree–Fock reference, CTSD and SC-CTSD are equivalent.

	$Be + He$[a]	$Be + 4He$[a]	$N_2 + He$[a]	$N_2 + 4He$[a]	$N_2 + H_2O$[b]
CISD	3.10	12.83	1.96	8.31	–
ACPF	0.56	0.92	0.41	1.11	0.88
AQCC	1.98	2.91	0.57	1.50	1.10
MRCI+Q	–	–	–	–	1.69
CASPT2	–	–	–	–	1.83
CASPT3	–	–	–	–	1.40
CTSD	0.00	0.00	0.00	0.00	0.00
SC-CTSD	0.00	0.00	0.00	0.00	0.00

[a] Reprinted with permission from Ref. [2], copyright 2007, American Institute of Physics. A Hartree–Fock reference was used.
[b] A CASSCF reference was used.

satisfy size consistency. Previous work [2] has investigated their size consistency errors when modeling the systems $Be + n He$ and $N_2 + n He$, in which all species are separated by 1000 Bohr. The size consistency error is then defined as the energy difference between treating the species separately or together. For simplicity, the Hartree–Fock solution was used as the reference wavefunction. As shown in Table 2, the size consistency errors of ACPF and AQCC varied from 0.5–0.9 mE_h and 2.0–2.9 mE_h, respectively, when modeling $Be + n He$. For $N_2 + n He$, the analogous errors ranged between 0.4–1.1 mE_h and 0.6–1.5 mE_h. In all cases, CT theory's size consistency error was precisely zero.

For this review, we have completed a short analysis of size consistency error in the system containing N_2 and H_2O separated by 1000 angstroms. In contrast to the previous study, we use a CASSCF reference to demonstrate size consistency in a multireference setting. We used the cc-pVDZ basis set [82], an N–N bond distance of 1.1208 angstroms, an O–H bond distance of 0.9929 angstroms, and an H–O–H bond angle of 109.57°. The CASSCF calculations optimized all orbitals using active spaces containing the H 1s, O 2p, and N 2p orbitals. During the subsequent dynamic correlation calculations, the O 1s and N 1s orbitals were not correlated. The resulting size consistency errors for CTSD using the exact overlap matrix and SC-CTSD were both less than 0.01 mE_h, showing that our implementation of CT theory is indeed size consistent when using a multi-configurational reference wavefunction. In contrast, MOLPRO [83] calculations on this system showed significant size consistency errors for CASPT2, CASPT3, ACPF, AQCC, and MRCI+Q, as shown in Table 2. We therefore see that among the methods available to treat dynamic correlation in multireference systems, CT theory is the only method that is strictly size consistent.

6.3. *Effect of truncation thresholds*

As described in Section 3.2, the standard CTSD method requires the choice of overlap matrix truncation thresholds τ_1 and τ_2 in order to prevent intruder states. Unfortunately, there is no way of knowing good values of these thresholds a priori, and in some systems good values may not exist at all. It is therefore important to analyze the effect of these thresholds on the CT energy to avoid creating artificially accurate or inaccurate results. In what we term a well behaved system, there is a range of values for both thresholds over which the CT energy changes little or not at all. We say that the energy of such a system is independent of the truncation thresholds. In some difficult systems, there may be no obvious choice for the thresholds (especially for τ_1), and one then has to investigate the effect of thresholds on the potential energy surface and report the results. In a recent application of CTSD to NiO [5], we found that there was no region of values for τ_1 over which the CT energy was stable. We therefore presented three potential energy curves for three different choices of τ_1 so that the reader could see the differences. In this case it happened that all three choices gave results with relative (i.e. non-parallelity) errors of less than 2 kcal/mol when compared to MRCI+Q, so although absolute energies were sensitive to the threshold, relative energies were not.

A typical approach for analyzing the threshold-sensitivity of the CTSD energy is shown in Figure 2, where we have plotted the energy of the N_2 dimer against both thresholds. The bond length used is 1.1208 angstroms, with other parameters the same as in Figure 1. We see that τ_2 has a region of stable energy between 0.01 and 0.002, below which the CT equations are too poorly conditioned to solve numerically. Thus we say that the double excitations necessary to describe dynamic correlation have all been included at a threshold of 0.01. For τ_1, there is a region of stability between 0.2 and 0.09, below which the energy fluctuates until the equations stop converging at 0.008. We are therefore justified in the choices $\tau_1 = 0.1$ and $\tau_2 = 0.01$ used in Section 6.1. Were we instead to choose τ_1 in the fluctuating region, say 0.03, we would introduce unnecessary error in our bonding curve due to the fact that the intruder states causing the fluctuations change

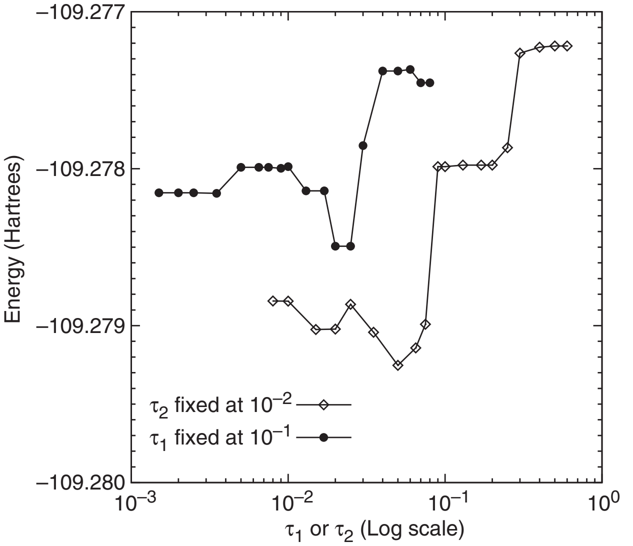

Figure 2. The dependence of the CTSD energy on the overlap truncation thresholds τ_1 and τ_2 for N_2 when using exact overlap matrices and a radius of 1.1208 angstroms. Other calculation parameters are the same as in Figure 1. Each curve represents the effect of varying one threshold while the other is held fixed.

with geometry. The necessity of validating the choice of τ_1 and τ_2 is bothersome in small systems and potentially unaffordable in large ones. In these cases it becomes desirable to first attempt the SC-CTSD method, whose energy is much less sensitive to changes in its single τ_ϵ threshold [5].

6.4. *Transition metal oxides*

One area in which CT theory can make important contributions is the quantitative study of transition metal chemistry. Transition metal electronic structure is in general difficult to treat with single-reference methods due to the strong correlations associated with the locality of the d orbitals. It also poses a significant problem for traditional multireference techniques due to the large number of valence orbitals involved. To explicitly correlate the 4s, 3d, and 4p orbitals of just two transition metals requires 18 active orbitals, which is already beyond the reach of CASSCF, CASPT2, and MRCI. Some interesting enzyme active sites, such as that found in aconitase, contain four or more transition metals, not to mention neighboring O and S atoms. Although advanced static correlation methods such as DMRG cannot yet treat such large active sites, they have been applied to $[Cu_2O_2]^{2+}$ clusters containing 32 active orbitals (see next section). CT theory is the only feasible method for treating dynamic correlation on top of such large active spaces. To assess the capabilities of CT theory for transition metal chemistry, we have studied CT theory's ability to treat single metal oxides, namely FeO and NiO, which are small enough to allow a comparison with traditional dynamic correlation methods.

In both FeO [2] and NiO [5], the CTSD method with exact overlap had a relative error of less than 2 kcal/mol when compared to MRCI+Q, as shown in Table 3. For comparison, CASPT2 had a 5.5 kcal/mol error in FeO and a 1.8 kcal/mol error in NiO. SC-CTSD, which unlike CTSD and CASPT2 does not use the three-body RDM, was also applied to NiO and produced a relative error of 3.7 kcal/mol. These results suggest that CTSD is more reliable in metal oxides than CASPT2 and give hope that it will be effective

Table 3. Results for the FeO and NiO potential energy curves. The MRCI+Q total energy is reported in E_h, with other methods reported as the difference from MRCI+Q in mE_h. The active space in both systems consists of the O 2p and metal 4s, 4p, and 3d orbitals. Data for FeO reprinted with permission from Ref. [2], copyright 2007, American Institute of Physics. Data for NiO reprinted with permission from Ref. [5], copyright 2010, American Institute of Physics. Relative errors are given with respect to MRCI+Q. See Section 6.4 for details.

FeO					NiO				
R[a]	MRCI+Q	CASSCF	CASPT2	CTSD[b]	R[a]	MRCI+Q	CASSCF	CASPT2	CTSD[c]
1.50	−1337.65843	299.22	−8.20	3.09	1.50	−1582.54099	751.29	−74.86	−3.11
1.65	−1337.67302	290.00	−5.73	3.00	1.60	−1582.55611	746.61	−73.25	−2.91
1.72	−1337.66923	284.82	−4.12	3.85	1.70	−1582.55476	742.20	−72.33	−2.79
2.00	−1337.62980	265.74	0.57	3.52	1.80	−1582.54261	736.47	−71.98	−5.20
Relative error (mE_h)		33.48	8.77	1.12			14.82	2.88	2.40

[a] Bond length in angstroms.
[b] $\tau_1 = 0.3$, $\tau_2 = 0.05$, exact overlap.
[c] $\tau_1 = 0.3$, $\tau_2 = 0.1$, exact overlap.

at describing larger transition metal clusters. While SC-CTSD was not as accurate as CTSD, it has an even greater potential to treat large clusters by avoiding the use of the three-body RDM. The study of NiO also revealed the importance of using the full set of valence orbitals, which becomes difficult when more than one transition metal is present. When the Ni 4p orbitals were excluded from the active space, the relative error of CASPT2 increased to 10.1 kcal/mol. In multi-metal clusters, methods such as CASPT2 and CTSD which require the three-body RDM will typically be forced to use an active space smaller than the full valence space and will likely suffer similar accuracy penalties. Another advantage of CT theory is its relatively low cost scaling, which in metal oxides begins to have a significant effect. The CASPT2 and CTSD calculations on FeO required a comparable amount of computation time, while MRCI+Q was more than an order of magnitude more expensive. In NiO, the MRCI+Q calculation was more than 50 times slower than SC-CTSD. In summary, CT theory has proven both affordable and accurate in initial studies on transition metal systems.

6.5. *CT and DMRG*

The combination of DMRG (for static correlation) and CT (for dynamic correlation) has recently been applied to achieve quantitative accuracy in a number of interesting systems whose active spaces are too large to be treated with traditional multireference methods. This section contains summaries of three such applications. For a more comprehensive analysis, we refer the reader to the original publications: Ref. [6] for polyenes and $[Cu_2O_2]^{2+}$ and Ref. [5] for free base porphin.

- *Electron correlation in polyenes*. In polyenes, the coupled cluster singles and doubles theory with perturbative triples, CCSD(T), is considered to give an excellent description of the ground states' correlation energies. Recent work has therefore used CCSD(T) as a benchmark for the performance of the multireference dynamic correlation methods CASPT2, MRACPF, and DMRG-SCF-CT, which used the full π orbital valence space as the active space. The results show that in the 6-31G basis set, DMRG-SCF-CT is the most accurate of the multireference methods, reproducing the CCSD(T) correlation energies to within 0.7 percent for polyenes ranging in length from four to 24 carbons. For comparison, MRACPF and CASPT2 correlation energies deviated by up to 1.8 and 12.1 percent, respectively, for polyenes between four and eight carbons long, the latter due to an insufficient description of dynamic correlation. Comparisons for longer polyenes were infeasible for MRACPF and CASPT2 due to their inability to model large active spaces. We therefore see that the combination of orbital optimized DMRG theory and CT theory can accurately model large systems that are only weakly multireference, which are usually difficult for multireference dynamic correlation theories due to both their large active spaces and the dominantly dynamic character of the electron correlation.
- *Copper oxide isomerization*. The isomerization of $[Cu_2O_2]^{2+}$ between its bis(μ-oxo) and μ-η^2:η^2 forms presents a particularly difficult challenge for *ab initio* methods due to the large variation in its multireference biradical character. While multireference methods should be well suited to treat this

changing nature in a balanced way, doing so in practice requires a much larger active space than can be treated with CASSCF and CASPT2. Indeed, all previous attempts to model the isomerization using CASPT2 produced *qualitatively* incorrect results in which an unphysical minimum appears midway through the isomerization [75,84–86]. DMRG and CT theory can afford to include in the active space all 32 necessary valence and double-shell orbitals [80] (the O 2p, O 3p, Cu 3d, and Cu 4d). With this larger active space, orbital-optimized DMRG (DMRG-SCF) succeeds in restoring qualitative accuracy. The SC-CTSD method is then applied to this reference state to produce a quantitative model of the isomerization curve which we believe to be accurate, although definitive experimental comparisons are unavailable. In addition to the unprecedented active space, this study also employs a larger basis set than any previous study of bare $[Cu_2O_2]^{2+}$. The application to the $[Cu_2O_2]^{2+}$ isomerization is an example of the potential of the joint DMRG and CT theory to tackle complex problems in transition metal chemistry.

- *Free base porphin*. Porphyrins play important roles in many biological processes such as oxygen absorption and photosynthesis, making their excitation spectrums of great interest. As for the $[Cu_2O_2]^{2+}$ complex, using multireference methods to model porphyrins in a balanced way is challenging due to their large numbers of valence orbitals. The combination of DMRG-SCF and SC-CTSD attempts to overcome this obstacle and has recently been used to compute the singlet-triplet gap of free base porphin, which is the parent compound for other porphyrins. The active space consisted of 26 electrons distributed among the 24 out-of-plane 2p orbitals of C and N. The resulting gap of 1.95 eV is higher than the 1.58 eV experimental value [87]. The exact source of this discrepancy remains to be explored, although some portion must be due to the fact that the SC-CTSD energy refers to a vertical transition while the experimental value includes the relaxation of the excited state geometry. Interestingly, unlike an earlier CASPT2 study in a 14 orbital active space which suffered from intruder states [88], no intruder states were encountered in SC-CTSD. Like $[Cu_2O_2]^{2+}$, free base porphin is an example of how CT and DMRG can be combined to model multireference systems with active spaces too large to be approached by other methods.

7. Conclusions and future directions

Canonical transformation (CT) theory is a dynamic correlation theory formulated for efficient application to multireference systems. As we have seen, the theory achieves many of the desirable properties that one looks for in a multireference dynamic correlation theory, including size consistency, the ability to treat large active spaces, and an n^6 cost scaling proportional to that of CCSD. In addition to these good formal properties, CT theory has demonstrated good performance in both benchmark systems, such as N_2, FeO, and NiO, and systems at the limits of multireference quantum chemical description, such as $[Cu_2O_2]^{2+}$ and free base porphin, with a typical accuracy significantly exceeding that of multireference perturbation theory. CT theory is a relatively young theory, however, and there remain a number of unanswered questions and possible improvements, which we shall now discuss.

(1) There is currently no implementation of CT theory available to the quantum chemistry community, a situation we hope to remedy soon by releasing an independent program package containing the CT and DMRG-SCF/CASSCF methods. Our intention is for this package to interface with existing quantum chemistry packages such as MOLPRO [83] and PSI3 [89] for the purpose of importing integrals, orbital coefficients, and the like.

(2) Two improvements that can be made to a dynamic correlation theory are the use of explicit correlation and density fitting techniques. While we foresee no fundamental difficulties in augmenting the theory with an explicit correlation method, the use of density fitting is not so straightforward. Initial investigations on the N_2 molecule have revealed that the CT effective Hamiltonian may lack the considerable two-body integral sparsity that is typically exploited by density fitting techniques. It is therefore not obvious how to employ density fitting in CT theory, but we believe the topic to be of significant research interest because it may simplify the challenge of constructing and storing the effective Hamiltonian.

(3) Another common feature of dynamic correlation theories is the ability to evaluate analytic derivatives. We are currently developing this capability for CT theory, which will allow the evaluation of nuclear gradients as well as response properties like polarizabilities.

(4) A less well understood but no less interesting topic is whether the canonical transformation can be optimized *simultaneously* with the underlying static correlation wavefunction in order to better capture the effect of active space relaxations that result from the inclusion of dynamic correlation. This simultaneous procedure would also eliminate the dichotomy between the 'diagonalize and perturb' and 'perturb and diagonalize' approaches to multireference models. However, little is understood as to how the cumulant and operator decompositions would effect the stability of such an optimization. As we have seen, great care is required to prevent these approximations from undermining the stability of CT theory itself, so it is natural to be cautious when entangling them with a static correlation method.

(5) Although CT theory is primarily intended for application to large multireference systems that are out of reach of traditional methods, it has proven remarkably competitive with existing methods in small molecules such as N_2 and H_2O. It would therefore be prudent to formulate a version of the theory that can make use of not only the exact three-body RDM but also a pre-contracted four-body RDM of the sort that is commonly employed in multireference perturbation theory. This would create a more accurate and robust method for small systems, and perhaps even large systems with small active spaces. Along similar lines, it would be interesting to develop a single-reference version of CT theory, in which many simplifications and efficiency gains should be possible. Research in these directions would allow theorists to select the type of canonical transformation theory most suitable to their specific application. It would also allow for more direct and informative comparisons between the behaviour of CT theory, single-reference coupled cluster

theories, and multireference theories that make use of higher order wavefunction information.

(6) Another area in which CT theory should be more extensively studied is the modeling of excited states. Our preliminary investigations into this topic have produced promising results but have also raised a number of questions about how best to approach the problem. For example, should one canonical transformation be defined for all excited states, akin to the similarity transform in equation-of-motion coupled cluster theory, or should each excited state be modeled with its own transformation that is specifically tailored to its reference wavefunction? In addition to answering such questions, the more basic question of how well CT theory performs for excited states has yet to be systematically investigated. In short, more extensive studies of excited states using CT theory are called for.

(7) Finally, we would like to raise the question of whether there is a better way to formulate efficient models for multireference dynamic correlation than through the use of internally contracted excitations. In this review we have advocated the use of even more restricted formulations, such as strong contraction, that increase the stability of CT theory. However, these excitations are still ultimately based on the internally contracted excitations employed in perturbation theory and configuration interaction. The usual justification for using internal contraction is that the obvious alternative of considering excitations from individual determinants in the active space becomes impractical for large active spaces, as the number of such determinants grows exponentially. However, the advent of advanced wavefunctions such as DMRG's matrix product state allow the active space wavefunction to be expressed in forms that require only a polynomial amount of information. It would be interesting to search for external excitations that are more suitable for use with the matrix product state and other advanced wavefunctions, for which the 'natural' starting point is something other than the uncontracted excitations that arise from a CASSCF wavefunction. Such excitations could have important implications for CT theory, but also for the general problem of modeling dynamic correlation in multireference systems.

Acknowledgements

This work was supported by the National Science Foundation CAREER program CHE-0645380, the David and Lucile Packard Foundation, the Alfred P. Sloan Foundation, and the Camille and Henry Dreyfus Foundation. E.N. would like to acknowledge the support of the National Science Foundation Graduate Research Fellowship Program.

References

[1] T. Yanai and G. K.-L. Chan, J. Chem. Phys. **124**, 194106 (2006).
[2] T. Yanai and G. K.-L. Chan, J. Chem. Phys. **127**, 104107 (2007).
[3] G. K.-L. Chan and T. Yanai, Adv. Chem. Phys. **134**, 343 (2007).
[4] E. Neuscamman, T. Yanai, and G. K.-L. Chan, J. Chem. Phys. **130**, 124102 (2009).
[5] E. Neuscamman, T. Yanai, and G. K.-L. Chan, J. Chem. Phys. **132**, 024106 (2010).
[6] T. Yanai, Y. Kurashige, E. Neuscamman, and G. K.-L. Chan, J. Chem. Phys. **132**, 024105 (2010).

[7] B. O. Roos, Adv. Chem. Phys. **69**, 399 (1987).
[8] K. Ruedenberg, M. W. Schmidt, M. M. Gilbert, and S. T. Elbert, Chem. Phys. **71**, 41 (1982).
[9] A. Szabo and N. Ostlund, *Modern Quantum Chemistry* (McGraw-Hill, New York, 1989).
[10] S. R. Langhoff and E. R. Davidson, Int. J. Quantum Chem. **8**, 61 (1974).
[11] E. R. Davidson and D. W. Silver, Chem. Phys. Lett. **52**, 403 (1977).
[12] R. Ahlrichs, P. Scharf, and C. Ehrhard, J. Chem. Phys. **82**, 890 (1985).
[13] R. J. Bartlett and M. Musial, Rev. Mod. Phys. **79**, 291 (2007).
[14] K. Andersson, P.-Å. Malmqvist, B. O. Roos, A. J. Sadlej, and K. Wolinski, J. Phys. Chem. **94**, 5483 (1990).
[15] K. Hirao, Chem. Phys. Lett. **190**, 374 (1992).
[16] C. Angeli, R. Cimiraglia, S. Evangelisti, T. Leininger, and J.-P. Malrieu, J. Chem. Phys. **114**, 10252 (2001).
[17] C. Angeli, R. Cimiraglia, and J.-P. Malrieu, J. Chem. Phys. **117**, 9138 (2002).
[18] R. J. Buenker and S. D. Peyerimhoff, Theor. Chim. Acta **35**, 33 (1974).
[19] P. E. M. Siegbahn, J. Chem. Phys. **72**, 1647 (1980).
[20] H.-J. Werner and P. J. Knowles, J. Chem. Phys. **89**, 5803 (1988).
[21] R. J. Gdanitz and R. Ahlrichs, Chem. Phys. Lett. **143**, 413 (1988).
[22] P. G. Szalay and R. J. Bartlett, Chem. Phys. Lett. **214**, 481 (1993).
[23] J. Paldus and X. Li, Adv. Chem. Phys. **110**, 1 (1999).
[24] S. R. White, J. Chem. Phys **117**, 7472 (2002).
[25] W. Kutzelnigg, J. Chem. Phys. **77**, 3081 (1982).
[26] W. Kutzelnigg, J. Chem. Phys. **80**, 822 (1984).
[27] R. J. Bartlett, S. A. Kucharski, and J. Noga, Chem. Phys. Lett. **155**, 133 (1989).
[28] J. D. Watts, G. W. Trucks, and R. J. Bartlett, Chem. Phys. Lett. **157**, 359 (1989).
[29] A. G. Taube and R. J. Bartlett, Int. J. Quantum Chem. **106**, 3393 (2006).
[30] S. Pal, M. D. Prasad, and D. Mukherjee, Theor. Chim. Acta **62**, 523 (1983).
[31] S. Pal, Theor. Chim. Acta **66**, 207 (1984).
[32] A. Banerjee and J. Simons, Int. J. Quantum. Chem. **14**, 207 (1981).
[33] A. Banerjee and J. Simons, J. Chem. Phys. **76**, 4548 (1982).
[34] M. R. Hoffmann and J. Simons, J. Chem. Phys. **88**, 993 (1988).
[35] D. Mukherjee, in *Recent Progress in Many-Body Theories*, edited by E. Schachinger, Vol. 4 (Plenum, New York, 1995), p. 127.
[36] D. Mukherjee, Chem. Phys. Lett. **274**, 561 (1997).
[37] W. Kutzelnigg and D. Mukherjee, J. Chem. Phys. **107**, 432 (1997).
[38] F. Colmenero and C. Valdemoro, Phys. Rev. A. **47**, 979 (1993).
[39] H. Nakatsuji and K. Yasuda, Phys. Rev. Lett. **76**, 1039 (1996).
[40] D. A. Mazziotti, Phys. Rev. A. **57**, 4219 (1998).
[41] W. Kutzelnigg and D. Mukherjee, J. Chem. Phys. **110**, 2800 (1999).
[42] D. Mukherjee and W. Kutzelnigg, J. Chem. Phys. **114**, 2047 (2001).
[43] D. A. Mazziotti, Phys. Rev. Lett. **97**, 143002 (2006).
[44] K. F. Freed, in *Many-Body Methods in Quantum Chemistry*, edited by U. Kaldor (Springer, Berlin, 1989), p. 1.
[45] B. Kirtman, J. Chem. Phys. **75**, 798 (1981).
[46] M. R. Hoffmann, Chem. Phys. Lett. **210**, 193 (1993).
[47] H. Nakatsuji, Chem. Phys. Lett. **59**, 362 (1978).
[48] J. Geertsen, M. Rittby, and R. J. Bartlett, Chem. Phys. Lett. **164**, 57 (1989).
[49] W. Kutzelnigg, Chem. Phys. Lett. **64**, 383 (1979).
[50] L. Cohen and C. Frishberg, Phys. Rev. A. **13**, 927 (1976).
[51] H. Nakatsuji, Phys. Rev. A. **14**, 41 (1976).
[52] F. E. Harris, Int. J. Quantum Chem. **90**, 105 (2002).
[53] J. M. Herbert, Int. J. Quantum Chem. **107**, 703 (2006).

[54] R. J. Bartlett, S. A. Kucharski, J. Noga, J. D. Watts, and G. W. Trucks, in *Many-Body Methods in Quantum Chemistry*, edited by U. Kaldor (Springer, Berlin, 1989), p. 125.
[55] N. Forsberg and P.-Å. Malmqvist, Chem. Phys. Lett. **274**, 196 (1997).
[56] K. G. Dyall, J. Chem. Phys. **102**, 4909 (1995).
[57] G. K.-L. Chan and M. Head-Gordon, J. Chem. Phys. **116**, 4462 (2002).
[58] G. K.-L. Chan and M. Head-Gordon, J. Chem. Phys. **118**, 8551 (2003).
[59] G. K.-L. Chan, J. Chem. Phys. **120**, 3172 (2004).
[60] G. K.-L. Chan, M. Kállay, and J. Gauss, J. Chem. Phys. **121**, 6110 (2004).
[61] G. K.-L. Chan and T. Van Voorhis, J. Chem. Phys. **122**, 204101 (2005).
[62] J. Hachmann, W. Cardoen, and G. K.-L. Chan, J. Chem. Phys. **125**, 144101 (2006).
[63] J. Hachmann, J. J. Dorando, M. Avilés, and G. K.-L. Chan, J. Chem. Phys. **127**, 134309 (2007).
[64] J. J. Dorando, J. Hachmann, and G. K.-L. Chan, J. Chem. Phys. **127**, 084109 (2007).
[65] D. Ghosh, J. Hachmann, T. Yanai, and G. K.-L. Chan, J. Chem. Phys. **128**, 144117 (2008).
[66] G. K.-L. Chan, Phys. Chem. Chem. Phys. **10**, 3454 (2008).
[67] S. R. White and R. L. Martin, J. Chem. Phys. **110**, 4127 (1999).
[68] S. Daul, I. Ciofini, C. Daul, and S. R. White, Int. J. Quantum Chem. **79**, 331 (2000).
[69] A. O. Mitrushenkov, G. Fano, F. Ortolani, R. Linguerri, and P. Palmieri, J. Chem. Phys. **115**, 6815 (2001).
[70] A. O. Mitrushenkov, R. Linguerri, P. Palmieri, and G. Fano, J. Chem. Phys. **119**, 4148 (2003).
[71] Ö. Legeza, J. Röder, and B. A. Hess, Phys. Rev. B **67**, 125114 (2003).
[72] Ö. Legeza, J. Röder, and B. A. Hess, Mol. Phys. **101**, 2019 (2003).
[73] G. Moritz, B. A. Hess, and M. Reiher, J. Chem. Phys. **122**, 024107 (2005).
[74] G. Moritz, A. Wolf, and M. Reiher, J. Chem. Phys. **123**, 184105 (2005).
[75] K. H. Marti, I. M. Ondík, G. Moritz, and M. Reiher, J. Chem. Phys. **128**, 014104 (2008).
[76] D. Zgid and M. Nooijen, J. Chem. Phys. **128**, 014107 (2008).
[77] D. Zgid and M. Nooijen, J. Chem. Phys. **128**, 144116 (2008).
[78] D. Zgid and M. Nooijen, J. Chem. Phys. **128**, 144115 (2008).
[79] J. J. Dorando, J. Hachmann, and G. K.-L. Chan, J. Chem. Phys. **130**, 184111 (2009).
[80] Y. Kurashige and T. Yanai, J. Chem. Phys. **130**, 234114 (2009).
[81] D. Ghosh, E. Neuscamman, T. Yanai, and G. K.-L. Chan, J. Chem. Theory Comput. (2009), submitted.
[82] T. H. Dunning, J. Chem. Phys. **90**, 1007 (1989).
[83] H.-J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, and M. Schütz, MOLPRO, version 2006, a package of *ab initio* programs, see http://www.molpro.net.
[84] C. J. Cramer, A. Kinal, M. Włoch, P. Piecuch, and L. Gagliardi, J. Phys. Chem. A. **110**, 11557 (2006).
[85] M. Flock and K. Pierloot, J. Phys. Chem. A. **103**, 95 (1999).
[86] M. F. Rode and H.-J. Werner, Theor. Chem. Acc. **114**, 309 (2005).
[87] M. Gouterman and G.-E. Khalil, J. Mol. Spectrosc. **53**, 88 (1974).
[88] L. Serrano-Andrés, M. Merchán, M. Rubio, and B. O. Roos, Chem. Phys. Lett. **295**, 195 (1998).
[89] T. D. Crawford, C. D. Sherill, E. F. Valeev, J. T. Fermann, R. A. King, M. L. Leininger, S. T. Brown, C. L. Janssen, E. T. Seidl, J. P. Kenny, and W. D. Allen, J. Comput. Chem. **28**, 1610 (2007).
[90] H. Larsen, J. Olsen, P. Jørgensen, and O. Christiansen, J. Chem. Phys. **113**, 6677 (2000).

A. Appendix

In this appendix we provide the explicit tensor contraction formulae necessary to evaluate the approximate commutator $[H, A]_{1,2}$ and amplitude equations $\langle\Psi_0|[H, \hat{o}_i]_{1,2}|\Psi_0\rangle$. The appendix is organized into three sections. First, we provide explicit definitions of the Hamiltonian, amplitude operator, and cumulant-approximated three-body RDM in the spin free form. Second, we present the tensor equations necessary for the approximate commutator. Finally, we present the tensor equations for the amplitude equations.

Note that in this appendix, repeated indices are assumed to be summed over, as per the Einstein convention. We use the index labels p and q for general indices, a for indices that are core or active, and e for indices that are active or virtual.

A.1. Spin free form of Hamiltonian, amplitudes, cumulant decomposition of density matrices

The spin free form of the Hamiltonian is given by

$$H = h_0 + h_1 + h_2, \tag{45a}$$

$$h_1 = t^{p_1}_{p_2} \hat{E}^{p_1}_{p_2}, \tag{45b}$$

$$h_2 = \frac{1}{2} v^{p_1 p_2}_{p_3 p_4} \hat{E}^{p_1 p_2}_{p_3 p_4}, \tag{45c}$$

in which h_0 is a constant and $\hat{E}$ are the unitary group generators defined in Equation (36). The spin free form of the amplitude operator A that defines the canonical transformation is

$$A = A_1 + A_2, \tag{46a}$$

$$A_1 = A^{e_1}_{a_1} (\hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}), \tag{46b}$$

$$A_2 = \frac{1}{2} A^{e_1 e_2}_{a_1 a_2} (\hat{E}^{e_1 e_2}_{a_1 a_2} - \hat{E}^{a_1 a_2}_{e_1 e_2}). \tag{46c}$$

Note that regardless of whether we use overlap truncation or strong contraction to define the set of operators $\{\hat{o}_i\}$, the amplitude operator A may be written in this general form as $\{\hat{o}_i\}$ themselves are linear combinations of the spin free excitation operators $(\hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1})$ and $(\hat{E}^{e_1 e_2}_{a_1 a_2} - \hat{E}^{a_1 a_2}_{e_1 e_2})$. Finally, the cumulant-approximated spin free three-body RDM is given by

$$\begin{aligned} \breve{D}^{p_1 p_2 p_3}_{q_1 q_2 q_3} = \langle \Psi_0 | (\hat{E}^{p_1 p_2 p_3}_{q_1 q_2 q_3})_{1,2} | \Psi_0 \rangle = {} & D^{p_1}_{q_1} \breve{D}^{p_2 p_3}_{q_2 q_3} - \frac{1}{2} D^{p_1}_{q_3} \breve{D}^{p_3 p_2}_{q_1 q_2} - \frac{1}{2} D^{p_1}_{q_2} \breve{D}^{p_2 p_3}_{q_1 q_3} \\ & + D^{p_2}_{q_2} \breve{D}^{p_3 p_1}_{q_3 q_1} - \frac{1}{2} D^{p_2}_{q_1} \breve{D}^{p_1 p_3}_{q_2 q_3} - \frac{1}{2} D^{p_2}_{q_3} \breve{D}^{p_3 p_1}_{q_2 q_1} \\ & + D^{p_3}_{q_3} \breve{D}^{p_1 p_2}_{q_1 q_2} - \frac{1}{2} D^{p_3}_{q_2} \breve{D}^{p_2 p_1}_{q_3 q_1} - \frac{1}{2} D^{p_3}_{q_1} \breve{D}^{p_1 p_2}_{q_3 q_2}, \end{aligned} \tag{47}$$

where

$$\breve{D}^{p_1 p_2}_{q_1 q_2} = D^{p_1 p_2}_{q_1 q_2} - \frac{2}{3} (D^{p_1}_{q_1} D^{p_2}_{q_2} - \frac{1}{2} D^{p_1}_{q_2} D^{p_2}_{q_1}). \tag{48}$$

A.2. Formulas for commutators

The commutator $[H, A]_{1,2}$ takes the following general form:

$$[H, A]_{1,2} = C_0 + C_1{}^{p_1}_{q_1} \hat{E}^{p_1}_{q_1} + C_2{}^{p_1 p_2}_{q_1 q_2} \hat{E}^{p_1 p_2}_{q_1 q_2}. \tag{49}$$

By separating H and A into their one- and two-body components, we may also write the approximate commutator as

$$[H, A]_{1,2} = [h_1, A_1]_{1,2} + [h_1, A_2]_{1,2} + [h_2, A_1]_{1,2} + [h_2, A_2]_{1,2}. \tag{50}$$

Thus, the coefficients C_0, ${C_1}^{p_1}_{q_1}$, and ${C_2}^{p_1p_2}_{q_1q_2}$ are each decomposed into the contributions from these four components as follows,

$$\begin{aligned} C_0 &= c_0, \\ {C_1}^{p_1}_{q_1} &= {c_1}^{p_1}_{q_1} + {c'_1}^{p_1}_{q_1}, \\ {C_2}^{p_1p_2}_{q_1q_2} &= {c_2}^{p_1p_2}_{q_1q_2} + {c'_2}^{p_1p_2}_{q_1q_2} + {c''_2}^{p_1p_2}_{q_1q_2}. \end{aligned} \tag{51}$$

The expressions to evaluate the above decomposed elements, c_0, c_1, c'_1, c_2, c'_2, and c''_2, are shown below.

We begin by formulating the simplest commutator $[h_1, A_1]_{1,2}$, which is expressed as,

$$[h_1, A_1]_{1,2} = h^{p_1}_{q_1} A^{e_1}_{a_1} \left[\hat{E}^{p_1}_{q_1}, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}\right]_{1,2} = {c_1}^{p_1}_{q_1} \hat{E}^{p_1}_{q_1}, \tag{52}$$

where the matrix ${c_1}^{p_1}_{q_1}$ is given by the symmetrization of the matrix ${\bar{c}_1}^{p_1}_{q_1}$,

$${c_1}^{p_1}_{q_1} = \frac{1}{2}\left({\bar{c}_1}^{p_1}_{q_1} + {\bar{c}_1}^{q_1}_{p_1}\right), \tag{53}$$

and the matrix elements ${\bar{c}_1}^{p_1}_{q_1}$ are determined from the following matrix products,

$${\bar{c}_1}^{p_1}_{q_1} = \left(h^{p_1}_{e_1} A^{e_1}_{a_1}\right)\delta^{q_1}_{a_1} - \left(h^{p_1}_{a_1} A^{e_1}_{a_1}\right)\delta^{q_1}_{e_1}. \tag{54}$$

Next, the commutator $[h_1, A_2]_{1,2}$ is written as

$$[h_1, A_2]_{1,2} = \frac{1}{2} h^{p_1}_{q_1} A^{e_1e_2}_{a_1a_2} \left[\hat{E}^{p_1}_{q_1}, \hat{E}^{e_1e_2}_{a_1a_2} - \hat{E}^{a_1a_2}_{e_1e_2}\right]_{1,2} = {c_2}^{p_1p_2}_{q_1q_2} \hat{E}^{p_1p_2}_{q_1q_2}, \tag{55}$$

where the symmetrization for the four-index array ${c_2}^{p_1p_2}_{q_1q_2}$ is carried out as follows,

$${c_2}^{p_1p_2}_{q_1q_2} = \frac{1}{4}\left({\bar{c}_2}^{p_1p_2}_{q_1q_2} + {\bar{c}_2}^{p_2p_1}_{q_2q_1} + {\bar{c}_2}^{q_1q_2}_{p_1p_2} + {\bar{c}_2}^{q_2q_1}_{p_2p_1}\right), \tag{56}$$

and the formula to calculate the array elements ${\bar{c}_2}^{p_1p_2}_{q_1q_2}$ is given by

$${\bar{c}_2}^{p_1p_2}_{q_1q_2} = \left(\frac{1}{2} h^{p_1}_{e_1} A^{e_1e_2}_{a_1a_2}\right)\delta^{p_2}_{e_2}\delta^{q_1}_{a_1}\delta^{q_2}_{a_2} - \left(\frac{1}{2} h^{p_1}_{a_1} A^{e_1e_2}_{a_1a_2}\right)\delta^{p_2}_{a_2}\delta^{q_1}_{e_1}\delta^{q_2}_{e_2}. \tag{57}$$

Finally, the commutators involving h_2 are described as

$$[h_2, A_1]_{1,2} = \frac{1}{2} v^{p_1p_2}_{q_1q_2} A^{e_1}_{a_1} \left[\hat{E}^{p_1p_2}_{q_1q_2}, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}\right]_{1,2} = {c'_2}^{p_1p_2}_{q_1q_2} \hat{E}^{p_1p_2}_{q_1q_2}, \tag{58}$$

$$[h_2, A_2]_{1,2} = \frac{1}{2} v^{p_1p_2}_{q_1q_2} A^{e_1e_2}_{a_1a_2} \left[\hat{E}^{p_1p_2}_{q_1q_2}, \hat{E}^{e_1e_2}_{a_1a_2} - \hat{E}^{a_1a_2}_{e_1e_2}\right]_{1,2} = c_0 + {c'_1}^{p_1}_{q_1} \hat{E}^{p_1}_{q_1} + {c''_2}^{p_1p_2}_{q_1q_2} \hat{E}^{p_1p_2}_{q_1q_2}. \tag{59}$$

The tensor product formulas to evaluate the coefficients c_0, $\bar{c}'_1$, ${\bar{c}'_2}^{p_1p_2}_{q_1q_2}$, and ${\bar{c}''_2}^{p_1p_2}_{q_1q_2}$ (to be symmetrized if needed) are shown below. In these formulas, we emphasize that a single index of each of the four-index arrays ${\bar{c}'_2}^{p_1p_2}_{q_1q_2}$, ${\bar{c}''_2}^{p_1p_2}_{q_1q_2}$, and $v^{p_1p_2}_{q_1q_2}$ is common. Denoting this common index as a bold index $\mathbf{p_1}$, we can calculate the output elements ${\bar{c}'_2}^{\mathbf{p_1}p_2}_{q_1q_2}$ and ${\bar{c}''_2}^{\mathbf{p_1}p_2}_{q_1q_2}$ ($\forall\ p_2, q_1, q_2$) from the input array $v^{\mathbf{p_1}p'_2}_{q'_1q'_2}$ ($\forall\ p'_2, q'_1, q'_2$) for a given $\mathbf{p_1}$, and thus the required storage in fast memory is $O(n^3)$. This trick is exploited in our implementation to achieve memory savings and a simplified parallelization. We now have

$${\bar{c}'_2}^{\mathbf{p_1}p_2}_{q_1q_2} = \left(\frac{1}{2} v^{\mathbf{p_1}p_2}_{e_1q_2} A^{e_1}_{a_1}\right)\delta^{q_1}_{a_1} - \left(\frac{1}{2} v^{\mathbf{p_1}p_2}_{a_1q_2} A^{e_1}_{a_1}\right)\delta^{q_1}_{e_1}, \tag{60}$$

$$c_0 = \left(v^{\mathbf{p_1}p_2}_{e_1q_2} A^{e_1e_2}_{a_1a_2}\right)\bar{D}^{\mathbf{p_1}p_2e_2}_{a_1q_2a_2} - \left(v^{\mathbf{p_1}p_2}_{a_1q_2} A^{e_1e_2}_{a_1a_2}\right)\bar{D}^{\mathbf{p_1}p_2a_2}_{e_1q_2e_2}, \tag{61}$$

where the definition of $\bar{D}^{p_1p_2p_3}_{q_1q_2q_3}$ is given from that of $\breve{D}^{p_1p_2p_3}_{q_1q_2q_3}$ (Equation 47) by replacing $\breve{D}^{p_1p_2}_{q_1q_2}$ (Equation 48) in it with $\bar{D}^{p_1p_2}_{q_1q_2}$, defined by

$$\bar{D}^{p_1p_2}_{q_1q_2} = -D^{p_1p_2}_{q_1q_2} + \frac{4}{3}\left(D^{p_1}_{q_1}D^{p_2}_{q_2} - \frac{1}{2}D^{p_1}_{q_2}D^{p_2}_{q_1}\right). \tag{62}$$

We may also evaluate

$$\bar{c}'^{\,\mathbf{p}_1}_{1q_1} = e'^{\,\mathbf{p}_1}_{1q_1} - a'^{\,\mathbf{p}_1}_{1q_1}, \tag{63}$$

in which various useful intermediates are defined as

$$e'^{\,\mathbf{p}_1}_{1q_1} = \left(v^{\mathbf{p}_1 i}_{q_1e_1} - \frac{1}{2}v^{\mathbf{p}_1 i}_{e_1q_1}\right){S_0}^{e_1}_{i} + \delta^{q_1}_{a_1}\left(v^{\mathbf{p}_1 i}_{e_1 j}\,{S_0}^{e_1 i}_{a_1 j} - v^{\mathbf{p}_1 i}_{j e_1}\,{S_1}^{e_1 i}_{a_1 j}\right)$$

$$- \delta^{q_1}_{e_2}\left(v^{\mathbf{p}_1 e_1}_{j i}\,{S_4}^{e_1e_2}_{ij}\right) + \delta^{\mathbf{p}_1}_{a_2}\delta^{q_1}_{e_2}\left(A^{e_1e_2}_{a_1a_2} - \frac{1}{2}A^{e_2e_1}_{a_1a_2}\right){S_6}^{a_1}_{e_1}, \tag{64}$$

$$a'^{\,\mathbf{p}_1}_{1q_1} = \left(v^{\mathbf{p}_1 i}_{q_1a_1} - \frac{1}{2}v^{\mathbf{p}_1 i}_{a_1q_1}\right){S_0}^{i}_{a_1} + \delta^{q_1}_{e_1}\left(v^{\mathbf{p}_1 i}_{a_1 j}\,{S_0}^{e_1 i}_{a_1 j} - v^{\mathbf{p}_1 i}_{j a_1}\,{S_1}^{e_1 i}_{a_1 j}\right)$$

$$- \delta^{q_1}_{a_2}\left(v^{\mathbf{p}_1 a_1}_{j i}\,{S_5}^{ij}_{a_1a_2}\right) + \delta^{\mathbf{p}_1}_{a_2}\delta^{q_1}_{e_2}\left(A^{e_1e_2}_{a_1a_2} - \frac{1}{2}A^{e_2e_1}_{a_1a_2}\right){S_6}^{e_1}_{a_1}, \tag{65}$$

$${S_0}^{e_1I}_{a_1J} = \left[A^{e_1e_2}_{a_1a_2} - \frac{1}{2}A^{e_1e_2}_{a_2a_1}\right]\tilde{D}^{e_2I}_{a_2J} \tag{66}$$

$${S_1}^{e_1I}_{a_1J} = \frac{1}{2}[A^{e_1e_2}_{a_2a_1}\tilde{D}^{e_2I}_{Ja_2}],\ + \frac{1}{2}[A^{e_1e_2}_{a_1a_2}\tilde{D}^{e_2I}_{a_2J}], \tag{67}$$

$${S_2}^{e_1}_{I} = [A^{e_1e_2}_{a_1a_2}\tilde{D}^{Ie_2}_{a_1a_2}], \tag{68}$$

$${S_3}^{I}_{a_1} = [A^{e_1e_2}_{a_1a_2}\tilde{D}^{Ia_2}_{e_1e_2}], \tag{69}$$

$${S_4}^{e_1e_2}_{IJ} = \frac{1}{2}[A^{e_1e_2}_{a_1a_2}\tilde{D}^{IJ}_{a_1a_2}], \tag{70}$$

$${S_5}^{IJ}_{a_1a_2} = \frac{1}{2}[A^{e_1e_2}_{a_1a_2}\tilde{D}^{IJ}_{e_1e_2}], \tag{71}$$

$${S_6}^{a_1}_{q_1} = [v^{p_1p_2}_{q_1q_2}\tilde{D}^{p_1p_2}_{a_1q_2}], \tag{72}$$

$$\tilde{D}^{p_1p_2}_{q_1q_2} = -D^{p_1p_2}_{q_1q_2} + 2\left(D^{p_1}_{q_1}D^{p_2}_{q_2} - \frac{1}{2}D^{p_1}_{q_2}D^{p_2}_{q_1}\right). \tag{73}$$

Note that the intermediates $S_0, S_1, \ldots, S_6$, and $\tilde{D}$ can be precomputed and stored in fast memory, since the size of the largest of them is $O(n^2_{\text{occ}}n^2_{\text{open}})$. Finally, we may evaluate

$$\bar{c}''^{\,\mathbf{p}_1p_2}_{2q_1q_2} = e''^{\,\mathbf{p}_1p_2}_{2q_1q_2} - a''^{\,\mathbf{p}_1p_2}_{2q_1q_2}, \tag{74}$$

using the intermediates

$$e''^{\,\mathbf{p}_1p_2}_{2q_1q_2} = \delta^{q_1}_{a_1}\delta^{q_2}_{a_2}\left(\frac{1}{4}v^{\mathbf{p}_1p_2}_{e_1e_2}A^{e_1e_2}_{a_1a_2}\right) \tag{75}$$

$$+ \delta^{p_2}_{e_2}\delta^{q_2}_{a_2}\left\{\left(v^{\mathbf{p}_1 i}_{q_1e_1} - \frac{1}{2}v^{\mathbf{p}_1 i}_{e_1q_1}\right)T^{(0)e_1e_2}{}_{ia_2} - \frac{1}{2}\left(v^{\mathbf{p}_1 i}_{q_1e_1}\,T^{(0)e_2e_1}{}_{ia_2}\right)\right\} - \delta^{p_2}_{e_2}\delta^{q_1}_{a_2}\left(\frac{1}{2}v^{\mathbf{p}_1 i}_{e_1q_2}\,T^{(0)e_2e_1}{}_{ia_2}\right) \tag{76}$$

$$- \delta^{q_1}_{a_1}\delta^{q_2}_{a_2}\left(\frac{1}{2}v^{\mathbf{p}_1p_2}_{e_1i}\,T^{(1)ie_1}{}_{a_1a_2}\right) + \delta^{q_1}_{a_1}\left(v^{\mathbf{p}_1p_2}_{e_1q_2}\,T^{(2)e_1}{}_{a_1}\right) + \delta^{q_1}_{a_1}\delta^{q_2}_{a_2}\delta^{p_2}_{e_2}\left(A^{e_1e_2}_{a_1a_2}\,T^{(3)\mathbf{p}_1}{}_{e_1}\right), \tag{77}$$

$$a''^{\,\mathbf{p}_1p_2}_{2q_1q_2} = \delta^{q_1}_{e_1}\delta^{q_2}_{e_2}\left(\frac{1}{4}v^{\mathbf{p}_1p_2}_{a_1a_2}A^{e_1e_2}_{a_1a_2}\right) \tag{78}$$

$$+ \delta^{p_2}_{a_2}\delta^{q_2}_{e_2}\left\{\left(v^{\mathbf{p}_1 i}_{q_1a_1} - \frac{1}{2}v^{\mathbf{p}_1 i}_{a_1q_1}\right)T^{(1)ie_2}{}_{a_1a_2} - \frac{1}{2}\left(v^{\mathbf{p}_1 i}_{q_1a_1}\,T^{(1)ie_2}{}_{a_2a_1}\right)\right\} - \delta^{p_2}_{a_2}\delta^{q_1}_{e_2}\left(\frac{1}{2}v^{\mathbf{p}_1 i}_{a_1q_2}\,T^{(1)ie_2}{}_{a_2a_1}\right) \tag{79}$$

$$- \delta^{q_1}_{e_1}\delta^{q_2}_{e_2}\left(\frac{1}{2}v^{\mathbf{p}_1p_2}_{a_1i}\,T^{(0)e_1e_2}{}_{ia_1}\right) + \delta^{q_1}_{e_1}\left(v^{\mathbf{p}_1p_2}_{a_1q_2}\,T^{(2)e_1}{}_{a_1}\right) + \delta^{q_1}_{e_1}\delta^{q_2}_{e_2}\delta^{p_2}_{a_2}\left(A^{e_1e_2}_{a_1a_2}\,T^{(3)\mathbf{p}_1}{}_{a_1}\right), \tag{80}$$

$$T_{0}{}^{e_1e_2}_{Ia_2} = [A^{e_1e_2}_{a_1a_2} D^{I}_{a_1}], \tag{81}$$

$$T_{1}{}^{Ie_2}_{a_1a_2} = [A^{e_1e_2}_{a_1a_2} D^{I}_{e_1}], \tag{82}$$

$$T_{2}{}^{e_1}_{a_1} = [A^{e_1e_2}_{a_1a_2} - \frac{1}{2} A^{e_2e_1}_{a_1a_2}] D^{e_2}_{a_2}, \tag{83}$$

$$T_{3}{}^{p_1}_{q_1} = [v^{p_1p_2}_{q_1q_2} - \frac{1}{2} v^{p_2p_1}_{q_1q_2}] D^{p_2}_{q_2}. \tag{84}$$

As before, the tensors $T_0, \ldots, T_3$ can be precomputed and stored in fast memory.

A.3. Formulas for residual elements

The residual elements are given by

$$R^{e_1}_{a_1} = \langle \Psi_0 | [\bar{H}, \hat{E}^{e_1}_{a_1}]_{1,2} | \Psi_0 \rangle, \tag{85a}$$

$$R^{e_1e_2}_{a_1a_2} = \langle \Psi_0 | [\bar{H}, \hat{E}^{e_1e_2}_{a_1a_2}]_{1,2} | \Psi_0 \rangle, \tag{85b}$$

and are decomposed into the contributions from the one- and two-body operators of $\bar{H}$,

$$R^{e_1}_{a_1} = R'{}^{e_1}_{1a_1} + R''{}^{e_1}_{1a_1}, \tag{86a}$$

$$R^{e_1e_2}_{a_1a_2} = R'{}^{e_1e_2}_{2a_1a_2} + R''{}^{e_1e_2}_{2a_1a_2}, \tag{86b}$$

where $R'{}^{e_1}_{1a_1}$ and $R'{}^{e_1e_2}_{2a_1a_2}$ are associated with the one-body operator h_1 (Equation 45b), and $R''{}^{e_1}_{1a_1}$ and $R''{}^{e_1e_2}_{2a_1a_2}$ with the two-body operator h_2 (Equation 45c). The tensor product forms of $R'{}^{e_1}_{1a_1}$ and $R'{}^{e_1e_2}_{2a_1a_2}$ are given by

$$R'{}^{e_1}_{1a_1} = \langle \Psi_0 | [h_1, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}]_{1,2} | \Psi_0 \rangle = 2(h^{p_1}_{e_1} D^{p_1}_{a_1} - h^{p_1}_{a_1} D^{p_1}_{e_1}), \tag{87}$$

$$R'{}^{e_1e_2}_{2a_1a_2} = \langle \Psi_0 | [h_1, \hat{E}^{e_1e_2}_{a_1a_2} - \hat{E}^{a_1a_2}_{e_1e_2}]_{1,2} | \Psi_0 \rangle = 2(h^{p_1}_{e_1} D^{p_1e_2}_{a_1a_2} - h^{p_1}_{a_1} D^{p_1a_2}_{e_1e_2}), \tag{88}$$

which can be easily implemented as the multiplications between the matrix $h^{p_1}_{q_1}$ and the one- and two-body RDMs, which are all kept in fast memory. The expressions to evaluate $R''{}^{e_1}_{1a_1}$ and $R''{}^{e_1e_2}_{2a_1a_2}$ are given by

$$R''{}^{e_1}_{1a_1} = \langle \Psi_0 | [h_2, \hat{E}^{e_1}_{a_1} - \hat{E}^{a_1}_{e_1}]_{1,2} | \Psi_0 \rangle = 2(v^{p_1p_2}_{e_1q_2} D^{p_1p_2}_{a_1q_2} - v^{p_1p_2}_{a_1q_2} D^{p_1p_2}_{e_1q_2}), \tag{89}$$

$$\begin{aligned} R''{}^{e_1e_2}_{2a_1a_2} &= \langle \Psi_0 | [h_2, \hat{E}^{e_1e_2}_{a_1a_2} - \hat{E}^{a_1a_2}_{e_1e_2}]_{1,2} | \Psi_0 \rangle \\ &= (v^{p_1p_2}_{e_1e_2} D^{p_1p_2}_{a_1a_2} - v^{p_1p_2}_{a_1a_2} D^{p_1p_2}_{e_1e_2}) + 2(v^{p_1p_2}_{e_1q_2} \breve{D}^{p_1p_2e_2}_{a_1q_2a_2} - v^{p_1p_2}_{a_1q_2} \breve{D}^{p_1p_2a_2}_{e_1q_2e_2}). \end{aligned} \tag{90}$$

The last two terms $v^{p_1p_2}_{e_1q_2} \breve{D}^{p_1p_2e_2}_{a_1q_2a_2}$ and $v^{p_1p_2}_{a_1q_2} \breve{D}^{p_1p_2a_2}_{e_1q_2e_2}$ are each seven-fold summations which at first glance have an $O(n^7)$ evaluation cost. This expense, however, can be reduced to $O(n^6)$ by using the decomposed form of the three-body RDM $\breve{D}^{p_1p_2e_2}_{a_1q_2a_2}$, given by Equation (47).