The Configuration Interaction Method: Advances in Highly Correlated Approaches

C. David Sherrill† and Henry F. Schaefer III

Center for Computational Quantum Chemistry
University of Georgia
Athens, Georgia 30602

Abstract

Highly correlated configuration interaction (CI) wavefunctions going beyond the simple singles and doubles (CISD) model space can provide very reliable potential energy surfaces, describe electronic excited states, and yield benchmark energies and molecular properties for use in calibrating more approximate methods. Unfortunately, such wavefunctions are also notoriously difficult to evaluate due to their extreme computational demands. The dimension of a full CI procedure, which represents the exact solution of the electronic Schrödinger equation for a fixed one-particle basis set, grows factorially with the number of electrons and basis functions. For very large configuration spaces, the number of CI coupling coefficients becomes prohibitively large to store on disk; these coefficients must be evaluated as needed in a so-called direct CI procedure. Work done by several groups since 1980 has focused on using Slater determinants rather than spin ($\hat{S}^2$) eigenfunctions because coupling coefficients are easier to compute with the former. We review the fundamentals of the configuration interaction method and discuss various determinant-based CI algorithms. Additionally, we consider some applications of highly correlated CI methods.

†Present address: Department of Chemistry, University of California, Berkeley, CA 94720-1460

ADVANCES IN QUANTUM CHEMISTRY, VOLUME 34

0065-3276/99 $30.00

Contents

1 Introduction

Most chemists picture the electronic structure of atoms or molecules by invoking orbitals. The orbital concept has its basis in Hartree-Fock theory, which determines the best wavefunction $|\Psi\rangle$ under the approximation that each electron experiences only the *average* field of the other electrons. This is also called the "one-electron," or "independent particle" model. While the Hartree-Fock method gives very useful results in many situations, it is not always quantitatively or even qualitatively correct. When this approximation fails, it becomes necessary to include the effects of electron correlation: one must model the *instantaneous* electron-electron repulsions present in the molecular Hamiltonian.

The most broadly applicable method for describing electron correlation is configuration interaction (CI), which expresses the wavefunction as a linear combination of $\hat{S}_z$ eigenfunctions (Slater determinants) or $\hat{S}_z$ and $\hat{S}^2$ eigenfunctions (configuration state functions, or CSFs) describing the distribution of N electrons. If all possible N-electron functions are included in the CI procedure (subject to spatial and spin symmetry restrictions), then the Schrödinger equation is solved exactly within the space spanned by the one-particle basis functions. Hence, in its most general form, CI applies to difficult cases such as excited states, open-shell systems, and systems far from their equilibrium geometries. However, the dimension of this "full CI" procedure grows factorially with molecular size, so it is necessary to select only the most important N-electron functions.

A common approach is to restrict the CI space to the Hartree-Fock self-consistent-field (SCF) configuration and those configurations related to it by single and double substitutions of orbitals, in a procedure denoted CISD. In cases where the SCF method yields a good approximate wavefunction, CISD with double-ζ plus polarization (DZP) single-particle basis sets typically predicts equilibrium bond lengths of small molecules within 0.4% of experiment and harmonic vibrational frequencies within 4%.[1] Unfortunately, CISD (and most other standard CI methods short of full CI) are not size extensive, meaning that their performance degrades with increasing molecular size. Size extensive alternatives include many-body perturbation theory (MBPT) and coupled-cluster (CC) methods. The coupled-cluster singles and doubles (CCSD) method outperforms CISD with only a moderately increased computational effort (the cost of both methods scales as the sixth power of the system size) because CCSD accounts for some triple and quadruple substitutions from the SCF configuration by approximating them as products of single and double substitutions. When employed with a triple-ζ plus double polarization basis set (TZ2P), the CCSD generally predicts bond lengths within 0.2% of experiment and harmonic vibrational frequencies within 2%.[2]

More accurate results can be obtained by using larger one-particle basis sets and employing the CCSD(T) method, which accounts for the effects of irreducible (or connected) triple substitutions in a single non-iterative step scaling as n^7.[3-5] Furthermore, recent equation-of-motion (EOM) or linear-response coupled-cluster theories for singly-excited electronic states[6,7] also outperform CISD in the prediction of excitation energies because CISD is biased towards the state described by the reference wavefunction.

Even the coupled-cluster methods eventually break down in cases where the SCF wavefunction is not a qualitatively correct description of the system. This can occur during bond-breaking reactions, for example, or for transition metals.[8,9] Hence, it is necessary to make the zeroth-order wavefunction multiconfigurational. Although multireference coupled-cluster theories are very difficult to formulate, multireference CI methods have been used for many years.[10,11] These methods typically include single and double substitutions from a set of "reference" configurations required to describe nondynamical electron correlation. Unfortunately, reference selection is not trivial, since the list of important references depends on the molecule and its geometry. This tends to make MR-CI methods unsuitable as a "model chemistry,"[12] since the quality of the wavefunction is not uniform across different molecules. One MR-CI wavefunction which is largely free from these difficulties is second-order CI (SOCI),[13] which is a multi-reference CISD in which the references are chosen as all possible distributions of electrons within a given "active space" (unfortunately, the acronym SOCI is also sometimes used to mean spin-orbit CI). The SOCI wavefunction requires much less computational effort than a full CI, yet it produces potential energy surfaces which nearly parallel the full CI surfaces.[14-17] If the active space is large enough, one can expect the SOCI method to provide equally good results for any small molecule at any geometry, making it a suitable model chemistry (SOCI is still not rigorously size extensive, so it may be necessary to apply size extensivity corrections for systems with eight electrons or more[15]). Unfortunately, the SOCI method is too computationally expensive to be generally applicable.

Hence there is a need to make SOCI more computationally efficient so that it can be used for larger chemical systems, and to develop related methods which scale better with the system size. Although the CI space can be reduced by individual selection of references or N-electron functions, for the reasons stated above it is beneficial to select the CI space in an *a priori* manner, once a minimal set of parameters, such as the active space, has been specified. For example, we have advocated a method we call CISD[TQ],[16-18] which is a SOCI in which higher-than-quadruple substitutions have been excluded. For systems dominated by a single reference, CISD[TQ] performs nearly as well as SOCI.[16,17]

Although extremely costly to determine, full CI results are invaluable in the calibration of such multireference CI methods, or indeed of essentially any *ab initio* electronic structure method, including many-body perturbation theory (MBPT)[19–23] and coupled-cluster approaches.[4,6,7,17,19,22,24–37] A few full CI benchmarks[17,38,39] have been carried out using the loop-driven graphical unitary group (LD-GUGA) CI approach,[40–43] which uses a spin eigenfunction, or CSF, basis. However, a majority of the full CI calculations to date have employed Slater determinants, even though this typically makes the CI vector 2-4 times longer. In 1980, Handy demonstrated that the benefits of Slater determinants can outweigh their disadvantages, primarily because the evaluation of the required matrix elements becomes so much simpler.[44] In his 1980 article, Handy introduced the alpha and beta string notation which has commonly been used in the development of new CI algorithms. After an important reformulation of the direct CI method by Siegbahn,[45] Knowles and Handy introduced a vectorized full CI algorithm that enabled a whole series of full CI benchmark studies by Bauschlicher, Langhoff, Taylor, and others.[15] In 1988, Olsen and co-workers showed how to improve the Knowles-Handy algorithm by reducing the operation count while still maintaining vectorization in the innermost loops.[46] Another important advance was the extension to certain types of truncated CI spaces in which determinants are chosen according to how many electrons they place in each of three orbital subspaces. This restricted active space (RAS) CI procedure is capable of evaluating SOCI and CISD[TQ] wavefunctions. Subsequently, other full CI algorithms involving basically the same amount of computational effort as Olsen's algorithm have been presented by Harrison and Zarrabian[47] and by Bendazzoli and Evangelisti.[48–50]

In this article, we provide an updated look at the configuration interaction method in general, and at highly correlated CI methods in particular. Special emphasis is given to methods which select the CI space in an *a priori* manner. After reviewing the basic theory of the CI method and the typical approximations employed, we discuss features common to all implementations: transformation of the one- and two-electron integrals from the atomic orbital to the molecular orbital basis, and iterative diagonalization methods. Next, we survey several determinant-based algorithms for full and RAS CI wavefunctions. We describe some technical issues in considerable detail and describe our experience with our own determinant-based CI program. Finally, we discuss some of the applications of highly correlated CI methods. Although crucial to the efficient determination of optimized geometries and vibrational frequencies, analytic gradients of CI wavefunctions are not discussed in this article; we refer the reader to the recent article by Shepard[51] and the monograph by Yamaguchi *et al.*[52] Furthermore, additional considerations may arise when designing a CI program to be used along with orbital optimization in the mul-

ticonfigurational (MC) or complete-active-space (CAS) SCF methods; these issues are discussed elsewhere.[9,53-56]

2 The Configuration Interaction Method

This section presents the essential elements of the configuration interaction method and is meant to be accessible to those who are not experts in CI. The classic review by Shavitt covers the theoretical fundamentals and various formulations given prior to 1977.[57] More recent reviews have been presented by Siegbahn,[58] Karwowski,[59] and Duch.[60]

2.1 Fundamentals

Configuration interaction is conceptually the simplest method for solving the time-independent electronic Schrödinger equation $\hat{H}|\Psi\rangle = E|\Psi\rangle$ under the Born-Oppenheimer approximation. The electronic wavefunction $|\Psi\rangle$ is approximated by a linear expansion of N-electron basis functions (where N is the number of electrons in the system), i.e.,

$$|\Psi\rangle = \sum_I c_I |\Phi_I\rangle. \tag{1}$$

The linear expansion coefficients c_I are the *CI coefficients*. Substituting this linear expansion into the electronic Schrödinger equation, one obtains[61] a matrix form more suitable for computation:

$$\mathbf{Hc} = E\mathbf{Sc}, \tag{2}$$

where the Hamiltonian operator $\hat{H}$ has been replaced by a matrix $\mathbf{H}$ and the CI wavefunction $|\Psi\rangle$ has been replaced by a column vector of coefficients $\mathbf{c}$. In principle, this "matrix mechanics" formulation is equivalent to the original electronic Schrödinger equation;[62] hence it is said that CI constitutes an "exact theory." In practice, however, the matrix equations are not exact because the expansion in equation (1) must be truncated to a finite number of terms. The matrix elements of the Hamiltonian are given by $H_{IJ} = \langle\Phi_I|\hat{H}|\Phi_J\rangle$ and $\mathbf{S}$ is the overlap matrix with elements $S_{IJ} = \langle\Phi_I|\Phi_J\rangle$. If orthonormal functions $|\Phi_I\rangle$ are used for the expansion, then of course $\mathbf{S}$ becomes the unit matrix and the equation becomes an eigenvalue equation. Since $\mathbf{H}$ is a Hermitian matrix, the number of orthogonal eigenvectors is equal to the dimension of the matrix. The lowest-energy solution represents the electronic ground state, and higher-energy solutions represent excited electronic states.

It is generally helpful to build into the expansion functions $\{|\Phi_I\rangle\}$ the symmetry properties of the system. According to the antisymmetry principle,

a wavefunction describing a system of electrons (more generally, fermions) must be antisymmetric with respect to the interchange of spatial and spin coordinates for any pair of electrons. This requirement is very commonly satisfied by making the expansion functions Slater determinants. A Slater determinant in which the one-electron functions $\phi_i, \phi_j, \ldots, \phi_k$ are occupied may be written

$$\Phi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_i(\mathbf{x}_1) & \phi_j(\mathbf{x}_1) & \ldots & \phi_k(\mathbf{x}_1) \\ \phi_i(\mathbf{x}_2) & \phi_j(\mathbf{x}_2) & \ldots & \phi_k(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \phi_i(\mathbf{x}_N) & \phi_j(\mathbf{x}_N) & \ldots & \phi_k(\mathbf{x}_N) \end{vmatrix} \tag{3}$$

and abbreviated as $|\phi_i \phi_j \ldots \phi_k\rangle$ or simply as $|ij \ldots k\rangle$. Note that this determinant is uniquely specified (up to a phase factor) by the list of occupied orbitals. It is easy to see that such a determinant satisfies the antisymmetry principle, since the interchange of coordinates for a pair of electrons translates to the swapping of rows of the determinant, which introduces a sign change. We may also write the above determinant as

$$\Phi = \frac{1}{\sqrt{N!}} \sum_P (-1)^p \hat{P} \phi_i(\mathbf{x}_1) \phi_i(\mathbf{x}_2) \cdots \phi_k(\mathbf{x}_N), \tag{4}$$

where $\hat{P}$ is a permutation of electron coordinates with sign $(-1)^p$. In this context it can be useful to define the *antisymmetrizer* as

$$\mathcal{A} = \frac{1}{\sqrt{N!}} \sum_P (-1)^p \hat{P}. \tag{5}$$

This operator produces a Slater determinant when applied to a simple product of spin orbitals. The antisymmetrizer is Hermitian, it commutes with $\hat{H}$, and its square is proportional to itself, i.e., $\mathcal{A}^2 = \sqrt{N!}\mathcal{A}$.

An electronic wavefunction can be described exactly by equation (1) if the expansion includes all possible Slater determinants formed from a complete set of one-electron functions $\{\phi\}$.[63] Such a procedure has been called *complete CI*.[40] Since a truly complete set of orbitals will typically be infinite, a complete CI is technically impossible to perform. However, if the one-electron basis set is truncated, then only a finite (albeit large) number of Slater determinants can be formed. Using all of these determinants in the expansion constitutes a *full CI* procedure, and the resulting eigenfunctions and eigenvalues are exact within the space spanned by the one-electron basis set. Although full CI results are extremely costly to compute, they are essential for benchmarking more approximate methods.

It is straightforward to show that if the exact wavefunction $|\Psi\rangle$ is an eigenfunction of some Hermitian operator $\hat{A}$, then expansion functions $|\Phi_I\rangle$ which are eigenfunctions of $\hat{A}$ with different eigenvalues do not contribute to the CI wavefunction and can be neglected in the expansion (1). If the Slater determinants are formed from spin-orbitals which are eigenfunctions of $\hat{s}_z$ and spatial symmetry operators, then the Slater determinants themselves will also be eigenfunctions of these spatial symmetry operators and of $\hat{S}_z$. However, a Slater determinant is not generally an eigenfunction of $\hat{S}^2$. Hence, a common alternative to Slater determinants are *configuration state functions* (CSFs), which are simply linear combinations of Slater determinants chosen to be eigenfunctions of $\hat{S}^2$. The benefit of using CSFs over determinants is that fewer N-electron functions are needed to describe the same state. The drawback is that matrix elements of the Hamiltonian are easier to compute using determinants. Of course, there are other possible choices for the N-electron basis functions. For instance, one can incorporate functions of two electrons (geminals),[64–67] as is done in the Hylleraas treatment of the helium atom.[68,69] Nevertheless, N-electron functions built from single-particle functions remain the most common.

Unfortunately, even with an incomplete one-electron basis, a full CI is computationally intractable for any but the smallest systems, due to the vast number of N-electron basis functions required (the size of the CI space is discussed in section 2.4.1). The CI space must be reduced, hopefully in such a way that the approximate CI wavefunction and energy are as close as possible to the exact values. By far the most common approximation is to begin with the Hartree-Fock procedure, which determines the best single-configuration approximation to the wavefunction that can be formed from a given basis set of one-electron orbitals (usually atom centered and hence called atomic orbitals, or AOs). This yields a set of molecular orbitals (MOs) which are linear combinations of the AOs:

$$\phi_i(\mathbf{x}_1) = \sum_{\mu} C_{\mu}^{i} \chi_{\mu}(\mathbf{x}_1), \tag{6}$$

where $\chi_\mu(\mathbf{x}_1)$ denotes an atomic orbital and C_μ^i is an *SCF coefficient.* The CI space can then be expanded according to substitution or "excitation" level relative to the SCF "reference" determinant, i.e.,

$$|\Psi\rangle = c_0|\Phi_0\rangle + \sum_{ia} c_i^a|\Phi_i^a\rangle + \sum_{a<b,i<j} c_{ij}^{ab}|\Phi_{ij}^{ab}\rangle + \sum_{a<b<c,i<j<k} c_{ijk}^{abc}|\Phi_{ijk}^{abc}\rangle + \ldots \tag{7}$$

where $|\Phi_i^a\rangle$ means the Slater determinant formed by replacing spin-orbital i in $|\Phi_0\rangle$ with spin orbital a, etc. The widely-employed CI singles and doubles (CISD) wavefunction includes only those N-electron basis functions which represent single or double substitutions relative to the reference state. Since the

Hamiltonian operator includes only one- and two-electron terms, only singly and doubly substituted configurations can interact directly with the reference, and they typically account for about 95% of the basis set correlation energy of small molecules at their equilibrium geometries,[38] where $|\Phi_0\rangle$ provides a good zeroth-order description. Truncation of the CI space according to excitation class is discussed more fully in section 2.4.1.

2.2 The Variational Theorem

One attractive feature of configuration interaction is that the computed lowest energy eigenvalue is always an upper bound to the exact ground state energy. This follows from the fact that the CI energy is given by the expectation value formula, or Rayleigh quotient,

$$E = \frac{\langle \Phi | \hat{H} | \Phi \rangle}{\langle \Phi | \Phi \rangle}. \tag{8}$$

The variational theorem may be proven by expressing the approximate wavefunction $|\Phi\rangle$ as a linear combination of the exact eigenvectors $|\Psi_i\rangle$; one easily obtains

$$E - \mathcal{E}_0 = \sum_i c_i^* c_i (\mathcal{E}_i - \mathcal{E}_0), \tag{9}$$

where $\mathcal{E}_i$ is the ith *exact* energy eigenvalue, i.e., $\hat{H}|\Psi_i\rangle = \mathcal{E}_i|\Psi_i\rangle$. Since the right-hand side of eq. (9) is necessarily non-negative, $E \geq \mathcal{E}_0$. Likewise, we can also insert an expansion over the exact eigenvectors *for a given one-electron space* to prove that the CI energy must be an upper bound to the full CI energy in the same one-electron basis set. Equation (9) demonstrates that the approximate wavefunction approaches the exact one ($c_0 \approx 1$) as the energy E is minimized (see section 2.2.1). Minimizing E is equivalent to minimizing the right-hand side of eq. (9); that is, the sum of squares of the absolute values of the coefficients of excited states is minimized with weight factors $(\mathcal{E}_i - \mathcal{E}_0)$. This means that other properties do not generally converge as quickly with CI space expansion as the energy. In fact, the error in the energy is quadratic in the wavefunction error. This can be shown by writing the energy as

$$E = \frac{\langle \Psi - \Delta\Psi | \hat{H} | \Psi - \Delta\Psi \rangle}{\langle \Psi - \Delta\Psi | \Psi - \Delta\Psi \rangle}, \tag{10}$$

with $|\Psi\rangle$ the exact wavefunction and the error $|\Delta\Psi\rangle$ chosen orthogonal to $|\Psi\rangle$. From this expression it is simple to demonstrate that all terms linear in $|\Delta\Psi\rangle$ are vanishing and that only quadratic terms remain.

It is easy to extend proofs of the variational theorem to the case of states which are the lowest roots of a given spatial and spin symmetry.[70] Since the

self-consistent-field (SCF) and multiconfigurational SCF (MCSCF) wavefunctions can be written as a linear expansion (1) containing one or a few Slater determinants, with an energy given by eq. (8), they also obey the variational theorem. Furthermore, just as the lowest CI eigenvalue is an upper bound to the exact ground-state energy, more generally any CI eigenvalue E_i is an upper bound to the corresponding exact excited state energy $\mathcal{E}_i$.[71] Additionally, as other N-electron basis functions are added to the CI space, the eigenvalues obey the MacDonald-Hylleraas-Undheim relations[71,72]

$$E_{i-1}^{(m)} \leq E_i^{(m+1)} \leq E_i^{(m)} \tag{11}$$

where m is the number of N-electron basis functions.

2.2.1 The Method of Linear Variations

Since the variational theorem proves that the energy of a CI wavefunction is always an upper bound to the exact energy, one might start simply from the linear expansion (1) and attempt to minimize the energy by varying the CI coefficients subject to the constraint that they remain normalized. It is easy to show[63] that this method of linear variations, or the Ritz method,[73] yields the matrix equation

$$\mathbf{Hc} = E\mathbf{Sc}. \tag{12}$$

That is, the method of linear variations is identical to the matrix formulation of the Schrödinger equation. Another way of viewing this result is that only solutions to eq. (12) are energetically stable with respect to variations in the linear expansion coefficients.

2.2.2 The Correlation Energy

Since the CI energy is always an upper bound to the exact energy, approximate CI methods can be judged according to what fraction of the correlation energy they recover. The correlation energy is defined as the difference between the energy in the Hartree-Fock limit (E_{HF}) and the exact nonrelativistic energy of a system ($\mathcal{E}_0$)

$$E_{\text{corr}} = \mathcal{E}_0 - E_{HF}. \tag{13}$$

This energy will always be negative because the Hartree-Fock energy is an upper bound to the exact energy. The exact nonrelativistic energy $\mathcal{E}_0$ could be calculated, in principle, via a full CI in a complete one-electron basis. If we have an incomplete one-electron basis set, then we can only compute the *basis set correlation energy*, which is the correlation energy for a given one-electron basis. Frequently the term correlation energy implies basis set correlation

energy. The correlation energy is the energy recovered by fully allowing the electrons to avoid each other; the Hartree-Fock method, rather than using the true instantaneous Coulomb repulsion between pairs of electrons, instead only allows each electron to experience an average potential due to all the other electrons. However, this description of *dynamical correlation* is not complete. When a molecule is pulled apart, the electrons should not need to avoid each other as much, so the magnitude of the correlation energy should decrease. In fact, the opposite is true, as shown by the basis set correlation energies given in Table 1 for H_2O at five different geometries.

Table 1: Correlation Energy in H_2O with a cc-pVDZ Basis as Both O–H Bonds are Stretched Simultaneously.

Geometry	E_{corr} (hartree)[a]
R_e	-0.217 821
$1.5 \cdot R_e$	-0.269 961
$2.0 \cdot R_e$	-0.363 954
$2.5 \cdot R_e$	-0.476 747
$3.0 \cdot R_e$	-0.567 554

[a]Data from Olsen *et al.*, ref 22.
All ten electrons are correlated.

The magnitude of the correlation energy increases as the O–H bonds are stretched beyond their equilibrium length because equation (13) also includes a more subtle effect called the *nondynamical* or *static* correlation energy. This part of the correlation energy reflects the inadequacy of a single reference in describing a given molecular state, and is due to nearly degenerate states or rearrangement of electrons within partially filled shells. Shavitt[57] has pointed out this deficiency in the correlation energy definition and has suggested that multiconfigurational Hartree-Fock may prove a more useful baseline than single-configuration Hartree-Fock in equation (13).

2.3 Matrix Elements in Terms of One- and Two-electron Integrals

2.3.1 Slater's Rules

The matrix elements $H_{IJ} = \langle \Phi_I | \hat{H} | \Phi_J \rangle$ can be expressed in terms of one- and two-electron integrals. If we employ Slater determinants, the matrix elements

may be evaluated using Slater's rules (also called the Slater-Condon rules)[74–76] if a common set of one-electron orbitals are used for all determinants and if these orbitals are orthonormal. If nonorthogonal orbitals are employed (e.g., atomic orbitals) then the more complicated Löwdin rules[77] apply.

Slater's rules are expressed here in terms of spin-orbitals, which are functions of the spatial and spin coordinates of a single electron. The one-electron integrals are written as

$$[i|\hat{h}|j] = \int \phi_i^*(\mathbf{x}_1)\hat{h}(\mathbf{x}_1)\phi_j(\mathbf{x}_1)d\mathbf{x}_1 \tag{14}$$

and the two-electron integrals, in Mulliken notation, are written as

$$[ij|kl] = \int \phi_i^*(\mathbf{x}_1)\phi_j(\mathbf{x}_1)\frac{1}{\mathbf{r}_{12}}\phi_k(\mathbf{x}_2)^*\phi_l(\mathbf{x}_2)d\mathbf{x}_1 d\mathbf{x}_2. \tag{15}$$

Before Slater's rules can be used, the two Slater determinants must be rearranged so that they have the maximum possible number of columns in common (recalling that each column swap causes a sign change). After the determinants are in maximum coincidence, we see how many spin orbitals they differ by and employ the following rules:

1. Identical Determinants:

$$\langle\Phi_1|\hat{H}|\Phi_1\rangle = \sum_m^N [i|\hat{h}|i] + \sum_{i>j}^N \{[ii|jj] - [ij|ji]\}\,. \tag{16}$$

2. Determinants that Differ by One Spin Orbital:

$$\begin{aligned} |\Phi_1\rangle &= |\cdots i\cdots\rangle \\ |\Phi_2\rangle &= |\cdots j\cdots\rangle \\ \langle\Phi_1|\hat{H}|\Phi_2\rangle &= [i|\hat{h}|j] + \sum_k^N \{[ij|kk] - [ik|kj]\}\,. \end{aligned} \tag{17}$$

3. Determinants that Differ by Two Spin Orbitals:

$$\begin{aligned} |\Phi_1\rangle &= |\cdots ij\cdots\rangle \\ |\Phi_2\rangle &= |\cdots kl\cdots\rangle \\ \langle\Phi_1|\hat{H}|\Phi_2\rangle &= [ik|jl] - [il|jk]. \end{aligned} \tag{18}$$

Some of the terms above may vanish after integrating over spin coordinates, and a pair of determinants differing by more than two spin orbitals have a matrix element of zero. A derivation of these rules can be found in the introductory text by Szabo and Ostlund.[63] The rules for evaluating Hamiltonian matrix elements in a CSF basis are more complicated and are generally derived[40,42,78] using second quantization, which we consider next.

2.3.2 Second Quantization

We are free to write the matrix elements in the more general form

$$H_{IJ} = \sum_{pq}^{n} \gamma_{pq}^{IJ} (p|\hat{h}|q) + \frac{1}{2} \sum_{pqrs}^{n} \Gamma_{pqrs}^{IJ} (pq|rs), \tag{19}$$

where the use of parentheses rather than square brackets denotes a switch to spatial orbital notation rather than spin orbital notation. Also note the factor of $1/2$ in the two-electron term. The constants γ_{pq}^{IJ} and Γ_{pqrs}^{IJ} are called the one- and two-electron coupling coefficients, respectively. The CI energy in terms of these coupling coefficients is

$$E = \sum_{IJ}^{CI} c_I \left[\sum_{pq}^{n} \gamma_{pq}^{IJ} (p|\hat{h}|q) + \frac{1}{2} \sum_{pqrs}^{n} \Gamma_{pqrs}^{IJ} (pq|rs) \right] c_J, \tag{20}$$

where we assume that the CI coefficients are real. The one- and two-electron reduced density matrices are defined as

$$\gamma_{pq} = \sum_{IJ}^{CI} c_I c_J \gamma_{pq}^{IJ}, \tag{21}$$

$$\Gamma_{pqrs} = \sum_{IJ}^{CI} c_I c_J \Gamma_{pqrs}^{IJ}, \tag{22}$$

and using these definitions the energy may be written more compactly as

$$E = \sum_{pq}^{n} \gamma_{pq} (p|\hat{h}|q) + \frac{1}{2} \sum_{pqrs}^{n} \Gamma_{pqrs} (pq|rs). \tag{23}$$

Some authors absorb the factor of $1/2$ into the definition of the two-electron coupling coefficient and reduced density matrix.

These coupling coefficients are generally derived using second quantization,[63,79] in which the Hamiltonian is written (for a given one-particle basis set) as

$$\hat{H} = \sum_{pq}^{2n} a_p^\dagger a_q [p|\hat{h}|q] + \frac{1}{2} \sum_{pqrs}^{2n} a_p^\dagger a_r^\dagger a_s a_q [pq|rs], \tag{24}$$

where $a_p^\dagger$ and a_p are the creation and annihilation operators, respectively, for an electron in spin orbital p. Note that the second-quantized form of the Hamiltonian is independent of the number of electrons. If the spatial parts of α and β spin orbitals are identical, it is easy to re-write the second-quantized Hamiltonian in terms of the following shift operators, which Paldus has shown[40] to be generators of the unitary group:

$$\hat{E}_{ij} = a_{i\alpha}^\dagger a_{j\alpha} + a_{i\beta}^\dagger a_{j\beta}. \tag{25}$$

Due to the anticommutation relations of creation and annihilation operators,

$$\hat{E}_{ij}^{\dagger} = \hat{E}_{ji} \tag{26}$$

$$[\hat{E}_{ij}, \hat{E}_{kl}] = \hat{E}_{il}\delta_{jk} - \hat{E}_{kj}\delta_{il}. \tag{27}$$

The resulting Hamiltonian in terms of these operators is

$$\hat{H} = \sum_{pq}^{n} h_{pq}\hat{E}_{pq} + \frac{1}{2}\sum_{pqrs}^{n}(pq|rs)\left(\hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps}\right), \tag{28}$$

where we have used the more compact notation $h_{pq} = (p|\hat{h}|q)$. It is clear that the one- and two-electron coupling coefficients can be written as

$$\gamma_{pq}^{IJ} = \langle\Phi_I|\hat{E}_{pq}|\Phi_J\rangle, \tag{29}$$

$$\Gamma_{pqrs}^{IJ} = \langle\Phi_I|\hat{E}_{pq}\hat{E}_{rs} - \delta_{qr}\hat{E}_{ps}|\Phi_J\rangle. \tag{30}$$

Furthermore, using equations (26) and (27), one can deduce the following:

$$\gamma_{pq}^{IJ} = (\gamma_{qp}^{JI})^* \tag{31}$$

$$\Gamma_{pqrs}^{IJ} = \Gamma_{rspq}^{IJ} = (\Gamma_{srqp}^{JI})^* = (\Gamma_{qpsr}^{JI})^*. \tag{32}$$

2.4 Reducing the Size of the CI Space

This section discusses strategies for reducing the number of N-electron basis functions in the CI space (given that, in the general case, it is impossible to include all of them). We have already discussed how N-electron functions with the wrong symmetry properties (e.g., point-group symmetry, or spin symmetry) can be dismissed immediately.

2.4.1 Truncating by Excitation Level

As noted in equation (7), the CI expansion is typically truncated according to excitation level; in the vast majority of CI studies, the expansion is truncated (for computational tractability) at doubly-substituted configurations. Since the Hamiltonian contains only two-body terms, only singles and doubles can interact directly with the reference; this is a direct result of Slater's rules (cf. section 2.3.1). Furthermore, the matrix elements of singly substituted determinants (or CSFs) with the reference are zero when canonical SCF orbitals are used, according to Brillouin's theorem. Hence, one expects double excitations to make the largest contributions to the CI wavefunction after the reference

state. Indeed, this is what is observed. Even though singles, triples, etc., do not interact directly with the reference, they can still become part of the CI wavefunction (i.e., have non-zero coefficients) because they mix with the doubles, directly or indirectly. Although singles are much less important to the energy than doubles, they are generally included in CI treatments because of their relatively small number and because of their greater importance in describing one-electron properties.

After singles and doubles, the most important determinants are triples and quadruples, because only these can interact directly with the doubles. The importance of a determinant to the final CI wavefunction is expected to fall off with increasing substitution or excitation level relative to the reference, assuming that the reference is a reasonable zeroth-order description of the desired electronic state. Table 2 demonstrates the importance of various excitation classes in obtaining CI energies. Singles and doubles account for 95% of the correlation energy at the equilibrium geometries of the molecules listed. Quadruple excitations are more important than triples, at least as far as the energy is concerned. At stretched geometries, the CISD and CISDT methods become markedly poorer, yet the CISDTQ method still recovers a very high (and nearly constant) fraction of the correlation energy, suggesting that CISDTQ should give reliable results for energy differences across potential energy surfaces for small molecules so long as no more than two bonds are broken at once (simultaneously breaking three bonds would require up to sextuple substitutions).

Table 3 demonstrates that the number of N-electron basis functions increases dramatically with increasing excitation level. A DZP basis should be considered the minimum adequate basis for a meaningful benchmark study.[15,81] While it is generally possible to perform CISD calculations on small molecules with a good one-electron basis, the CISDTQ method is limited to molecules containing very few heavy atoms, due to the extreme number of N-electron functions required. Full CI calculations are of course even more difficult to perform, so that despite their importance as benchmarks, few full CI energies using large one-electron basis sets have been obtained.

The size of the full CI space in CSFs can be calculated (including spin symmetry but ignoring spatial symmetry) by Weyl's dimension formula.[82] If N is the number of electrons, n is the number of orbitals, and S is the total spin, then the dimension of the CI space in CSFs is given by

$$D_{nNS} = \frac{2S+1}{n+1} \begin{pmatrix} n+1 \\ N/2-S \end{pmatrix} \begin{pmatrix} n+1 \\ N/2+S+1 \end{pmatrix}. \tag{33}$$

The dimension of the full CI space in determinants (again, ignoring spatial

Table 2: Percentage of correlation energy recovered by various CI excitation levels for some small molecules.

	Percent Corr. Energy[a]		
Molecule	CISD	CISDT	CISDTQ
BH	94.91	n/a	99.97
HF	95.41	96.49	99.86
H_7^+	96.36	96.87	99.96
$H_2O(R_e)$	94.48	95.85	99.85
$H_2O(1.5\ R_e)$	89.36	92.05	99.48
$H_2O(2.0\ R_e)$	80.21	84.59	98.40
NH_3	94.44	95.43	99.84

[a]Results are for a DZP basis and are taken from refs 38 (BH, HF), 17 (H_7^+), 22 (H_2O), and 80 (NH_3). H_2O results correlate all ten electrons and employ the cc-pVDZ basis.

Table 3: Number of CSFs required for small molecules at several levels of CI.

	CSFs required[a]			
Molecule	CISD	CISDT	CISDTQ	FCI
BH	568	n/a	28 698	132 686
HF	552	6 712	48 963	944 348
H_7^+	1 271	24 468	248 149	2 923 933
H_2O	1 311	27 026	332 491	94 165 610
NH_3	2 443	52 595	619 235	48 642 057

[a]Results are for a DZP basis and are taken from refs 38 (BH, HF), 17 (H_7^+), 22 (H_2O), and 80 (NH_3). H_2O results correlate all ten electrons and employ the cc-pVDZ basis.

Table 4: Dimension of Full CI in Determinants (CSFs in parentheses)

	Number of electrons			
Orbitals	6	8	10	12
10	14.4×10^3	44.1×10^3	63.5×10^3	44.1×10^3
	(4.95×10^3)	(13.9×10^3)	(19.4×10^3)	(13.9×10^3)
20	1.30×10^6	23.5×10^6	240×10^6	1.50×10^9
	(379×10^3)	(5.80×10^6)	(52.6×10^6)	(300×10^6)
30	16.5×10^6	751×10^6	20.3×10^9	353×10^9
	(4.56×10^6)	(172×10^6)	(4.04×10^9)	(62.5×10^9)

symmetry) is computed simply by

$$D_{nN_\alpha N_\beta} = \begin{pmatrix} n \\ N_\alpha \end{pmatrix} \begin{pmatrix} n \\ N_\beta \end{pmatrix}, \tag{34}$$

or, in a form closer to equation (33),

$$D_{nNS} = \begin{pmatrix} n \\ N/2+S \end{pmatrix} \begin{pmatrix} n \\ N/2-S \end{pmatrix}. \tag{35}$$

Table 4 shows the dimension of the full CI space (neglecting spatial symmetry) in determinants and in CSFs for closed-shell systems. Current full CI algorithms are typically limited to several million determinants. Although there have been reports of larger calculations (including more than a billion determinants[83,84]), the computational expense is currently too great for routine calculations of this size.

2.4.2 Multireference Configuration Interaction

A full CI wavefunction is invariant to orbital rotations and even to the choice of the reference function. By contrast, the simple CISD method is quite sensitive to the choice of reference and orbitals. This explains the poor performance of CISD when the bonds are stretched in H_2O (cf. Table 2): the SCF wavefunction becomes an inadequate reference at stretched geometries, and CISD is unable to overcome this inadequacy. Such difficulties can occur even at equilibrium geometries if multiple low-lying electronic states are present. For

example, the zeroth-order wavefunction for a singlet diradical often requires two electron configurations: one doubly occupies the MO formed from in-phase radical orbitals, while the other doubly occupies the out-of-phase MO. Another example is the $\tilde{c}$ state of CH_2, which also requires a two-configuration treatment: the two configurations correspond to the two choices for the lone pair of electrons, either in the molecular plane or perpendicular to it.[85,86] More than two configurations can be critical for transition metals or when multiple bonds are broken.

If a CISD procedure includes all the important N-electron functions from the zeroth-order wavefunction (the "references") and also the single and double substitutions for *each* of these references, then the resulting method is referred to as multireference (MR) CISD. The MR-CISD wavefunction may be written

$$|\Phi_{MRCI}\rangle = \sum_R c(R)|\Phi(R)\rangle + \sum_R \sum_{ix} c_i^x(R)|\Phi_i^x(R)\rangle + \sum_R \sum_{ijxy} c_{ij}^{xy}(R)|\Phi_{ij}^{xy}(R)\rangle, \tag{36}$$

where R denotes a reference function, and i, j (x, y) run over orbitals which are occupied (unoccupied) for a given R. Clearly, a determinant or CSF which is generated as a single or double substitution from one reference state might also be generated as a single or double from a different reference; only unique N-electron functions are included in the MR-CISD procedure. If a sufficient number of references are included, then a MR-CISD can provide results nearly as good as the full CI[14,15,87] at a dramatically reduced computational expense. In the MR-CI method, the set of orbitals which are occupied in any of the references constitutes the *internal space*, and all other orbitals are in the *external space*. Sometimes a further distinction is made among the internal space orbitals: those whose occupancy is constant for all references are called *inactive* (even though their electrons may be excited in the final wavefunction), and the rest are called *active*. In the direct CI method[88] (see section 4), it is more convenient to rewrite (36) in an equivalent form which emphasizes the number of external orbitals:

$$|\Phi_{MRCI}\rangle = \sum_I c_I|\Phi_I\rangle + \sum_S \sum_a c_S^a|\Phi_S^a\rangle + \sum_P \sum_{ab} c_P^{ab}|\Phi_P^{ab}\rangle, \tag{37}$$

where a and b are external orbitals, and I, S, and P denote internal states (including spin coupling) with N, $N-1$, and $N-2$ electrons, respectively.

One very straightforward, *a priori* selection scheme is to make references of all those N-electron functions which can be obtained by distributing electrons in all possible ways in a subset of the most important orbitals (the "active space"). This results in the second-order CI (SOCI),[13] which is known[16,17,89] to provide high-quality potential energy surfaces nearly parallel to those from a full CI. Unfortunately, this prescription typically produces too many references

and the final CI space is too large to be computationally tractable. One strategy which has received relatively little attention[17,46,90] is to approximate a SOCI by restricting the references according to their excitation level. It is often reasonable to assume that the most important references are single and double substitutions from a dominant single reference.[16,17,91] Making this restriction leads to a MR-CISD which includes those triples and quadruples which have no more than two electrons outside the active space. This wavefunction has been designated CISD[TQ] to emphasize the variational treatment of limited triples and quadruples, and it has been shown to closely match SOCI when a single reference configuration dominates.[16,17] Although the CISD[TQ] expansion is much smaller than SOCI, it remains intractable for systems with more than two or three heavy atoms. Further strategies to reduce the cost of a CISD[TQ] are discussed later.

A much more common procedure for reference selection is to accept references whose estimated importance is greater than some given threshold; this can involve perturbative estimates of a function's energetic contribution or its coefficient in some preliminary wavefunction. These approaches are more successful at obtaining the best wavefunction at the lowest expense, but they sacrifice the simplicity of the excitation class selection and can become more difficult to implement and to use. One complication is that potential energy surfaces determined using such methods may not be smooth; to alleviate this, one may need to determine the important references at each geometry and use the union of these sets at every point.

Discarding some of the single and double excitations is another way to reduce the CI space. As with reference selection, the most common approaches involve estimates of a function's energetic contribution or coefficient. The CIPSI method of Malrieu and co-workers selects determinants based on perturbation theory estimates of their coefficients in the first-order wavefunction.[92,93] Alternatively, Buenker and Peyerimhoff[10] select spatial orbital configurations on the basis of each configuration's energetic contribution to a small CI consisting only of the references and the CSFs formed from that configuration. Obviously this involves solving a very large number of small CI problems. Alternatively, one can estimate the importance of all configurations simultaneously via a procedure such as Gershgorn and Shavitt's B_k method.[94,95] Shavitt's 1977 review article[57] surveys these and related alternatives.

Finally, Siegbahn has suggested two procedures for reducing the number of variational parameters in a MR-CISD wavefunction. The first method, *externally contracted* MR-CISD,[88] expresses eq. (37) as

$$|\Phi_{EC-MRCI}\rangle = \sum_I c_I|\Phi_I\rangle + \sum_S c_S \sum_a \tilde{c}_S^a|\Phi_S^a\rangle + \sum_P c_P \sum_{ab} \tilde{c}_P^{ab}|\Phi_P^{ab}\rangle$$

$$= \sum_I c_I|\Phi_I\rangle + \sum_S c_S|\Phi_S\rangle + \sum_P c_P|\Phi_P\rangle, \tag{38}$$

where the external contraction coefficients $\tilde{c}_S^a$ and $\tilde{c}_P^{ab}$ are determined perturbatively and the coefficients c_I, c_S, and c_P are determined variationally; hence, this is a type of variational perturbation theory.[96-98] Note that the total number of variational coefficients is now drastically reduced compared to eq. (37). Siegbahn states[58] that the error in the correlation energy due to external contraction is roughly 1-3%.

An alternative contraction scheme which has received more attention is *internally contracted* multireference CISD (usually denoted simply CMRCI), which was first discussed by Meyer[99] and Siegbahn.[100] This method applies the single and double excitation operators to a single multiconfigurational reference wavefunction *as a whole*, including the reference coefficients. Thus, if the reference wavefunction is

$$|\Phi_0\rangle = \sum_R c_R|\Phi_R\rangle, \tag{39}$$

then there are at least three other classes of expansion functions—singly external, semi-internal, and doubly-external:[58]

$$|\Phi_i^a\rangle = \hat{E}_{ai}|\Phi_0\rangle = \sum_S d_S|\Phi_S^a\rangle \tag{40}$$

$$|\Phi_{ij}^{ak}\rangle = \hat{E}_{ai}\hat{E}_{kj}|\Phi_0\rangle = \sum_S d_S|\Phi_S^a\rangle \tag{41}$$

$$|\Phi_{ijp}^{ab}\rangle = (\hat{E}_{ai}\hat{E}_{bj} + p\hat{E}_{aj}\hat{E}_{bi})|\Phi_0\rangle = \sum_P d_P|\Phi_P^{ab}\rangle, \tag{42}$$

where p is +1 (-1) for singlet (triplet) coupling of a and b. The coefficients d are not variational parameters, but fixed linear combinations of the reference coefficients c_R. The final wavefunction is then

$$|\Phi_{IC-MRCI}\rangle = c_0|\Phi_0\rangle + \sum_{ia} c_i^a|\Phi_i^a\rangle + \sum_{ijka} c_{ij}^{ak}|\Phi_{ij}^{ak}\rangle + \sum_{ijpab} c_{ijp}^{ab}|\Phi_{ijp}^{ab}\rangle. \tag{43}$$

Once again, the contraction has dramatically reduced the number of variational parameters. One difficulty with the internally contracted multireference CI method is that the relevant coupling coefficients become considerably more difficult to calculate. Werner and Knowles[101] alleviate this problem by leaving internal and singly-external configurations uncontracted, i.e.,

$$|\Phi_{IC-MRCI}\rangle = \sum_I c_I|\Phi_I\rangle + \sum_{Sa} c_S^a|\Phi_S^a\rangle + \sum_{ijpab} c_{ijp}^{ab}|\Phi_{ijp}^{ab}\rangle. \tag{44}$$

The remaining coupling coefficients still require elements of the third and fourth order reduced density matrices, which can now be evaluated each time

they are required due to advances by Werner and Knowles.[54,102] The contraction error in the internally contracted MR-CISD method is generally only 0.1-0.2%.[58]

2.4.3 Other CI Selection Schemes

In 1988, Olsen and co-workers[46] presented the restricted active space (RAS) CI, which specifies the CI space in an *a priori* manner reminiscent of the second-order CI (SOCI) and its derivatives. Olsen partitions the orbitals into three subspaces, labeled RAS I, RAS II, and RAS III. Typically, RAS I contains occupied and possibly very important virtual orbitals, RAS II contains the most important virtuals, and RAS III contains the less important virtuals. The CI space includes all determinants with a minimum of p electrons in RAS I and a maximum of q electrons in RAS III. There is no restriction on RAS II, which is akin to the complete active space. Using this simple procedure, it is possible to formulate any CI space truncated according to excitation level (e.g., CISD, CISDT, etc.) as well as excitation class selected MR-CI spaces, such as SOCI and CISD[TQ]. The RAS CI method is discussed more fully in section 4.8.

There are of course many other possible ways to select the CI space. For example, it is possible to generalize the RAS scheme to allow for more flexible CI spaces; work along these lines is presented later. In contrast, it is also possible to make the CI selection process essentially random. A recent paper by Greer discusses the unusual strategy of selecting CSFs using a Monte Carlo algorithm.[103]

2.4.4 The First-Order Interacting Space

Another way of limiting the size of the CI space is to include only those N-electron functions which contribute to the first-order wavefunction in Rayleigh-Schrödinger perturbation theory. This is the motivation behind the *interacting space* classification: the zeroth-order interacting space consists of the set of references, and the first order interacting space includes all those N-electron functions which have a nonzero Hamiltonian matrix element with some member of the zeroth-order interacting space.[104–106] Similarly, it is possible to define n-th order interacting spaces as those functions having nonzero Hamiltonian matrix elements with some member of the $(n-1)$th order interacting space.[106] Various methods for obtaining the first-order interacting space have been presented by Bunge,[104] Schaefer and co-workers,[43,105] and McLean and Liu.[106]

The first-order interacting space restriction is generally used to reduce the number of double substitutions included in single- or multi-reference CISD

wavefunctions. Although single substitutions from a Hartree-Fock reference determinant may be noninteracting due to Brillouin's theorem, they are often included nevertheless because of their strong interaction with the doubles and their importance in describing one-electron properties. For closed shell systems, the first-order interacting space criterion is inconsequential because Slater's rules dictate that all double substitutions are interacting. However, for open-shell systems there can be CSFs whose *spatial* orbital configuration differs from the reference by two electrons but which are noninteracting because of their spin coupling. In a basis of Slater determinants, one can enforce the first-order interacting space restriction simply by ensuring that all determinants differ from the reference by at most two *spin orbitals*. Aside from being more compact, wavefunctions limited to the first-order interacting space can exhibit certain orbital invariance properties.[107]

2.4.5 Computational Scaling

Depending on the relative sizes of the number of electrons, the number of orbitals, and the excitation level, one can derive several different simple estimates of the computational cost of a configuration interaction procedure. Obviously that cost relates to the number of N-electron functions in the linear expansion of the wavefunction, and the size of the CI space for various methods has already been discussed in section 2.4.1.

For present purposes, it is sufficient to work with spin orbitals. Typically, the dimension of the CI space is dominated by determinants with the highest excitation level, m. Thus

$$N_{det} \sim \left(\begin{array}{c} N \\ m \end{array} \right) \left(\begin{array}{c} n_v \\ m \end{array} \right) \sim \frac{1}{(m!)^2} N^m n_v^m, \tag{45}$$

with n_v spin orbitals unoccupied in the reference. Most CI procedures solve only for the lowest or lowest few eigenvectors, via an iterative procedure (cf. section 3.2). In such situations, the scaling is much less than the $\mathcal{O}(N_{det}^3)$ typical of standard matrix diagonalization methods. The most expensive step in iterative procedures such as the Davidson method[108] is the construction of the so-called σ vectors,

$$\sigma_i = \mathbf{H}\mathbf{b}_i, \tag{46}$$

where $\mathbf{b}_i$ belongs to a set of trial vectors which is expanded each iteration until convergence is reached. If the Hamiltonian matrix $\mathbf{H}$ were formed directly, this procedure would require $\mathcal{O}(N_{det}^2)$ operations. This is never actually done because the storage requirements would be too great, and such an approach ignores the fact that the Hamiltonian contains only two-body terms, so that the majority of the matrix elements are zero.

Each element of a trial vector $\mathbf{b}_i$ need only be multiplied by the nonzero elements of $\mathbf{H}$. The Hamiltonian will connect a maximally excited determinant with other maximally excited determinants and with other determinants having excitation level $m \geq m' \geq m-2$. The number of interacting determinants is roughly

$$\begin{pmatrix} m \\ 2 \end{pmatrix} \begin{pmatrix} n_v + m \\ 2 \end{pmatrix} + \begin{pmatrix} m \\ 2 \end{pmatrix} \begin{pmatrix} N-m \\ 2 \end{pmatrix} + m^2 n_v (N-m), \tag{47}$$

which we further approximate as

$$\frac{1}{4} m^2 (n_v + m)^2 + \frac{1}{4} m^2 (N-m)^2 + m^2 n_v (N-m). \tag{48}$$

Each element in $\mathbf{b}_i$ must be multiplied by the relevant nonzero matrix elements, leading to an overall operation count on the order of

$$\mathcal{O}(N^m n_v^m \left\{ m^2 n_v^2 + m^2 N^2 + m^2 N n_v \right\}). \tag{49}$$

Except for full CI, we typically expect $N, n_v >> m$. Furthermore, we almost always have $n_v > N$, so that the leading term becomes

$$\mathcal{O}(N^m n_v^{2+m}). \tag{50}$$

Thus the number of operations for a CISD procedure has a sixth power dependence on the total number of orbitals, while CISDTQ scales as the tenth power. For a given system, the number of occupied orbitals is fixed, and the cost of increasing the basis set size scales as $\mathcal{O}(n_v^{2+m})$; for CISD and CISDTQ, this scaling becomes $\mathcal{O}(n_v^4)$ and $\mathcal{O}(n_v^6)$, respectively.

The scaling of a multireference CI procedure can be estimated by multiplying the single-reference scaling by the number of references. The scaling of the CISD[TQ] method, for example, is roughly $\mathcal{O}(N^4 n_v^4)$, since the number of references is roughly N^2 if the active space is small relative to the external orbital space. For very high levels of excitation, including full CI, the number of interacting matrix elements for a given determinant becomes approximately $N^2 n^2$, so that the computational cost becomes roughly

$$\mathcal{O}^{FCI}(N_{det} N^2 n^2), \tag{51}$$

where we have replaced the term $N^m n_v^m$ with the actual number of determinants, N_{det}. For comparison, for n spatial orbitals the determinant full CI algorithm of Knowles and Handy[109] scales as $\mathcal{O}(N_{det} n^4)$, while the algorithm of Olsen *et al.*[46] and similar approaches scale as $\mathcal{O}(N_{det} N^2 (n - N/2)^2)$ for a closed-shell system. Although the exponents appearing in (51) are smaller than those in (50), it is important to remember that N_{det} contains a factorial dependence on N and n (see section 2.4.1); hence, a full CI procedure is extremely demanding computationally.

2.4.6 Size Extensivity Corrections

If we truncate the CI (either in the one-electron or N-electron space), we no longer have an exact theory. Of course either of these truncations will introduce an error in the wavefunction, which will cause errors in the energy and all other properties. One particularly unwelcome result of truncating the N-electron basis is that CI energies are no longer size extensive or size consistent.

These two terms—size extensive and size consistent—are used somewhat loosely in the literature. Of the two, size extensivity is the most well-defined. A method is said to be size extensive if the energy calculated thereby scales linearly with the number of particles N; the word "extensive" is used in the same sense as in thermodynamics. A method is called size consistent if it gives an energy $E_A + E_B$ for two well separated subsystems A and B. While the definition of size extensivity applies at any geometry, the concept of size consistency applies only in the limiting case of infinite separation. In addition, size consistency *usually* also implies correct dissociation into fragments; this is the source of much of the confusion arising from this term. Thus restricted Hartree-Fock (RHF) is size extensive, but it is not necessarily size consistent, since it cannot properly describe dissociation into open-shell fragments. It can be shown that many-body perturbation theory (MBPT) and coupled-cluster (CC) methods are size extensive, but they will be size consistent only if they are based on reference wavefunction which dissociates properly.

As previously stated, truncated CI's are neither size extensive nor size consistent. A simple (and often used) example makes this clear. Consider two noninteracting hydrogen molecules. If the CISD method is used, then the energy of the two molecules at large separation will not be the same as the sum of their energies when calculated separately. For this to be the case, one would have to include *quadruple* excitations in the supermolecule calculation, since local double excitations could happen simultaneously on A and B.

Clearly the fraction of the correlation energy recovered by a truncated CI will diminish as the size of the system increases, making it a progressively less accurate method. There have been many attempts to correct the CI energy to make it size extensive. The most widely-used (and simplest) of these methods is referred to as the Davidson correction,[110,111] which is

$$\Delta E_{DC} = (1 - c_0^2)(E_{CISD} - E_{SCF}). \tag{52}$$

This correction approximately accounts for the effects of "unlinked quadruple" excitations (i.e. simultaneous pairs of double excitations), and there are many similar expressions in use. For instance, the "renormalized" Davidson correction[112] is

$$\Delta E_{RDC} = \frac{1 - c_0^2}{c_0^2}(E_{CISD} - E_{SCF}). \tag{53}$$

Note that, when $c_0^2 \approx 1$, the two versions are nearly equivalent. A number of other variations exist,[113,114] some of which force the correction to vanish for two-electron systems.

A multireference version of Davidson's correction is given by[115]

$$\Delta E_{MR-DC} = \left(1 - \sum_{i \in \text{Ref}} |c_i|^2\right)(E_{MRCI} - E_{MR}), \tag{54}$$

where E_{MRCI} is the multireference CI energy and E_{MR} is the energy obtained from a CI in the space spanned by the references. We have simply replaced the CISD correlation energy in eq. (52) with the analogous multireference correlation energy, and we have replaced c_0^2 with the analogous sum of squares of all the reference coefficients. If the sum of the squares of reference coefficients is not near unity, better results may be obtained by using the renormalized version of this equation:

$$\Delta E_{MR-RDC} = \frac{1 - \sum_{i \in \text{Ref}} |c_i|^2}{\sum_{i \in \text{Ref}} |c_i|^2}(E_{MRCI} - E_{MR}). \tag{55}$$

It should be noted, however, that for a fixed system size, increasing the number of references decreases size extensivity errors. Indeed, for very highly correlated MR-CI wavefunctions, applying corrections such as (54) and (55) can sometimes lead to less reliable results.

There are a number of other *a posteriori* size extensivity corrections, most of which are computationally trivial once the wavefunction has been obtained. Duch[114] compares several of the more common corrections. Of course it is also possible to allow coupling between the wavefunction and the size extensivity correction. This leads to such methods as the coupled electron pair approximation (CEPA),[116] and the coupled pair functional (CPF) approaches.[117] This is also the motivation behind the quadratic configuration interaction method of Pople, Head-Gordon, and Raghavachari.[118] These authors determine the correlation energy and CI coefficients for quadratic CI with singles and doubles (QCISD) by the following set of projection equations, in spin-orbital notation:

$$\langle \Phi_0 | \hat{H} | \hat{C}_2 \Phi_0 \rangle = E_{corr} \tag{56}$$

$$\langle \Phi_i^a | \hat{H} - E_{scf} | (\hat{C}_1 + \hat{C}_2 + \hat{C}_1 \hat{C}_2) \Phi_0 \rangle = c_i^a E_{corr} \tag{57}$$

$$\langle \Phi_{ij}^{ab} | \hat{H} - E_{scf} | (1 + \hat{C}_1 + \hat{C}_2 + \frac{1}{2}\hat{C}_2^2) \Phi_0 \rangle = c_{ij}^{ab} E_{corr}, \tag{58}$$

where intermediate normalization ($\langle \Phi_0 | \Phi \rangle = 1$) has been employed, the Brillouin condition has been assumed ($\langle \Phi_i^a | \hat{H} | \Phi_0 \rangle = 0$), and $\hat{C}_1$ and $\hat{C}_2$ are the

standard single and double substitution operators,

$$\hat{C}_1 = \sum_{ia} c_i^a a_a^\dagger a_i \tag{59}$$

$$\hat{C}_2 = \frac{1}{4} \sum_{ijab} c_{ij}^{ab} a_a^\dagger a_b^\dagger a_j a_i. \tag{60}$$

The QCISD projection equations differ from the equivalent CISD equations only in the addition of the quadratic terms $\hat{C}_1\hat{C}_2$ and $\frac{1}{2}\hat{C}_2^2$, which lead to size-extensive energies. Alternatively, the QCISD equations may be considered an approximation to CCSD in which certain terms have been neglected. Pople and co-workers show how to extend this approach to include triples fully (QCISDT) or perturbatively [QCISD(T)].[118]

A multireference method building approximate size extensivity into the wavefunction is the Averaged Coupled Pair Functional (ACPF) method of Gdanitz and Ahlrichs,[119] which introduces an electron number dependence into the denominator of the MR-CISD energy functional. A similar method has been presented by Szalay and Bartlett under the name multireference averaged quadratic coupled-cluster (MR-AQCC).[120,121] Also noteworthy is the work of Malrieu and co-workers, who have presented a state-specific self-consistent dressing of the MR-CISD Hamiltonian matrix which gives size extensive results.[122–124]

2.4.7 The Frozen Core Approximation

It is quite common in correlated methods (including many-body perturbation theory, coupled-cluster, etc., as well as configuration interaction) to invoke the frozen core approximation, whereby the lowest-lying molecular orbitals, occupied by the inner-shell electrons, are constrained to remain doubly-occupied in all configurations. The frozen core for atoms lithium to neon typically consists of the 1s atomic orbital, while that for atoms sodium to argon consists of the atomic orbitals 1s, 2s, $2p_x$, $2p_y$ and $2p_z$. The frozen molecular orbitals are those made primarily from these inner-shell atomic orbitals.

A justification for this approximation is that the inner-shell electrons of an atom are less sensitive to their environment than the valence electrons. Thus the error introduced by freezing the core orbitals is nearly constant for molecules containing the same types of atoms. In fact, it is often preferable to employ the frozen core approximation as a general rule because most of the basis sets commonly used in *ab initio* quantum chemistry do not provide sufficient flexibility in the core region to accurately describe the correlation of the core electrons. Recently, Woon and Dunning have attempted to alleviate this problem by publishing correlation consistent core-valence basis sets.[125]

Not only does the frozen core approximation reduce the number of configurations, but it also reduces the computational effort required to evaluate matrix elements between the configurations which remain. Assuming that all frozen core orbitals are doubly occupied and orthogonal to all other molecular orbitals, it can be shown[126] that

$$\langle \Phi_I | \hat{H} | \Phi_J \rangle = \langle \bar{\Phi}_I | \hat{H}_0 | \bar{\Phi}_J \rangle, \tag{61}$$

where $\bar{\Phi}_I$ and $\bar{\Phi}_J$ are identical to Φ_I and Φ_J, respectively, except that the core orbitals have been deleted from $\bar{\Phi}_I$ and $\bar{\Phi}_J$, and $\hat{H}$ has been replaced by $\hat{H}_0$ defined by

$$\hat{H}_0 = E_c + \sum_{i=1}^{N-N_c} \hat{h}_c(i) + \sum_{i>j}^{N-N_c} \frac{1}{r_{ij}}, \tag{62}$$

where N is the number of electrons and N_c is the number of core electrons. E_c is the so-called "frozen-core energy," which is the expectation value of the determinant formed from only the N_c core electrons doubly occupying the $n_c = N_c/2$ core orbitals

$$E_c = 2\sum_{i}^{n_c} h_{ii} + \sum_{ij}^{n_c} \{2(ii|jj) - (ij|ji)\}. \tag{63}$$

Finally, $\hat{h}_c(i)$ is the one-electron Hamiltonian operator for electron i in the average field produced by the N_c core electrons,

$$\hat{h}_c(i) = \hat{h}(i) + \sum_{j=1}^{n_c} \left\{ 2\hat{J}_j(i) - \hat{K}_j(i) \right\}, \tag{64}$$

with $\hat{J}_j(i)$ and $\hat{K}_j(i)$ representing the standard Coulomb and exchange operators, respectively. Note that, although we have written the frozen core energy E_c and frozen core operator $\hat{h}_c$ in terms of molecular orbitals, it is not necessary to explicitly transform the one- and two-electron integrals involving core orbitals. Assuming real orbitals, we can define a frozen core density matrix[127] in atomic (or symmetry adapted) orbitals as

$$P^c_{\rho\sigma} = \sum_{i}^{n_c} C^i_\rho C^i_\sigma, \tag{65}$$

where C^i_ρ is the contribution of atomic orbital ρ to molecular orbital i. Now the frozen core operator in atomic orbitals becomes

$$h^c_{\mu\nu} = h_{\mu\nu} + 2\sum_{\rho\sigma}(\rho\sigma|\mu\nu)P^c_{\rho\sigma} - \sum_{\rho\sigma}(\rho\mu|\nu\sigma)P^c_{\rho\sigma}, \tag{66}$$

and the frozen core operator in molecular orbitals h^c_{ij} can be obtained simply by transforming $h^c_{\mu\nu}$. Similarly the frozen core energy can be evaluated as

$$\begin{aligned} E_c &= \sum_{\mu\nu} P^c_{\mu\nu} \left(h_{\mu\nu} + h^c_{\mu\nu} \right) \\ &= Tr(P^c h) + Tr(P^c h^c). \end{aligned} \tag{67}$$

An analogous approximation is the *deleted virtual* approximation, whereby a few of the highest-lying virtual (unoccupied) molecular orbitals are constrained to remain unoccupied in all configurations. Since these orbitals can never be occupied, they can be removed from the CI procedure entirely. The rationalization for this procedure is that it is unlikely that electrons will choose to partially populate high-energy orbitals in their attempt to avoid other electrons. However, this conclusion is generally true only for very high-lying virtual orbitals (such as those formed by antisymmetric combinations of core orbitals for a given atom). For all other virtual orbitals, such simplistic reasoning is insufficient. Indeed, Davidson points out that those high energy SCF virtual orbitals which result from the antisymmetric combination of the two basis functions describing each valence atomic orbital in a double-ζ basis set (such as the 3p-like orbital formed from the minus combination of the larger and smaller 2p atomic orbitals on oxygen) often make the largest contribution to the correlation energy in Møller-Plesset (MPn) wavefunctions.[128]

2.5 Choice of Orbitals

The results of any configuration interaction procedure depend on the choice of the atomic orbital (AO) basis. However, for a fixed AO basis, certain choices of molecular orbitals give equivalent CI wavefunctions. CI wavefunctions which are based on a single closed-shell reference and are truncated at a given excitation level are invariant to nonsingular linear transformations which mix doubly occupied orbitals with each other or unoccupied orbitals with each other. The invariance properties of CI wavefunctions based on open-shell references are more complicated, and the energy is generally not invariant to the rotation of open-shell orbitals unless certain extra references are added or the spin couplings are restricted to the first-order interacting space.[107] A full CI is invariant to all nonsingular linear transformations among the orbitals, even those that mix occupied and unoccupied orbitals; hence, the choice of the "reference" is irrelevant for a full CI procedure.

Some of the more elaborate CI spaces also exhibit invariance properties. Shavitt has defined the *full class CI* as one which partitions the orbitals into an arbitrary number of orthogonal subsets and includes all or none of the

N-electron functions which have a given partitioning of electrons among the subspaces.[57] The RAS CI wavefunctions are of this type, as are the RAS formulations of SOCI and CISD[TQ]. Such CI wavefunctions are invariant to separate, nonsingular linear transformations within any of the orbital subspaces. This property is relevant to the formulation of analytic gradients.[129]

For bond breaking processes, the restricted Hartree-Fock approximation will not yield a good reference. This can be remedied by employing a generalized valence bond (GVB)[130] reference or an unrestricted Hartree-Fock reference. However, the latter entails spin contamination in the CI wavefunction by states of higher spin multiplicity. Another alternative is to use a multiconfigurational (MC)[53] or complete-active-space (CAS) SCF[55,131] reference, which can be constructed to behave properly at all locations on the potential energy surface.

For multireference CI's such as SOCI and CISD[TQ], or any RAS CI which uses the RAS II orbital subspace, it is important that the orbitals of the active space be good correlating orbitals (i.e., they should be localized in the same region of space as the occupied orbitals). This is equally important for selected CI procedures, in that the number of configurations needed to achieve a given accuracy will be reduced. This localization criterion is not generally satisfied by canonical SCF virtual orbitals, whose construction is not physically motivated because they are based on an N-electron potential rather than an $(N-1)$ electron potential. One possible solution is to determine the virtual orbitals using a different (and more suitable) effective Hamiltonian than that used for the occupied orbitals, and to orthogonalize the resulting orbitals against the occupied orbitals. This is the procedure in the improved virtual orbital (IVO) method of Hunt and Goddard.[132] IVO's look like excited state orbitals and are more contracted than canonical SCF orbitals. Nevertheless, they remain somewhat too diffuse for making the CI expansion as small as possible. A related and improved method is the modified virtual orbital (MVO) approach of Bauschlicher,[133] who obtains virtual orbitals by diagonalizing the virtual subspace of a Fock matrix constructed for the core electrons only (eq. 64). Another possibility for obtaining compact virtual orbitals is Davidson's K-orbital approach.[134]

More commonly, good correlating orbitals are obtained with the MCSCF[53] or CASSCF[55] methods. Yet another possibility are the natural orbitals from a CISD wavefunction. Natural orbitals (NOs)[77,135] are defined as the eigenfunctions of the one-particle density matrix; the eigenvalues are called the occupation numbers of the NOs. One drawback of NOs is that the Hamiltonian is no longer diagonally-dominant,[136] and this can decrease the efficiency of iterative diagonalization methods (section 3.2).

Grev and Schaefer have shown for a number of small molecules that the

SOCI method performs just as well when based on CISD NOs as when based on CASSCF orbitals.[16] Furthermore, they demonstrated that this is not merely due to the highly accurate treatment of correlation in the SOCI method, because a SOCI wavefunction based on canonical SCF orbitals performs notably worse. One advantage of CISD natural orbitals is that they can be easier to compute than CASSCF orbitals. Another is that their importance to the CI wavefunction falls off very rapidly with occupation number. This means that one can delete several of the most weakly-occupied NOs from the MR-CISD procedure with little loss in the correlation energy recovered; such considerations do not necessarily hold for high-lying MCSCF or CASSCF orbitals. Additionally, Parisel and Ellinger have investigated the use of CI natural orbitals in variation-perturbation methods which employ a CI wavefunction as the zeroth-order solution in a subsequent second-order perturbation treatment,[137] and Blomberg and Liu have considered the use of CI natural orbitals in SOCI transition moment calculations.[138] Balasubramanian uses SOCI natural orbitals in his relativistic CI procedure.[139]

Finally, it has long been recognized that spatially localized orbitals should allow an efficient truncation of the CI space (see, for example, the PCILO method of Malrieu and co-workers[140,141]). SCF orbitals can be localized according to the Boys procedure[142] or various other methods. In most cases, the savings realized should outweigh any extra effort due to the loss of point-group symmetry. In the 1980s, Saebø and Pulay developed various localized correlation methods (including CISD) which can achieve computational savings in two distinct ways: first, the pair correlation energy for distant pairs can be neglected (or estimated), and second, the set of virtual orbitals used as correlating orbitals can be restricted to the atomic orbitals in the vicinity of the orbital to be correlated (with components of occupied orbitals projected out).[143,144] Since standard CISD scales as the sixth power of the system size, some type of localized correlation treatment is inevitable as quantum chemists seek to apply correlated methods to large molecules.

2.6 Excited Electronic States

Here we will briefly discuss configuration interaction descriptions of excited electronic states. As previously mentioned, excited states are described by higher-energy eigenvectors of the Hamiltonian. However, since one can apply spin and spatial symmetry restrictions to the N-electron basis functions, solving for excited states which are energetically the lowest of a given symmetry species proceeds just as for the ground state. In this way, one can use orbitals which are optimal for each state.

Much more challenging is the case when several states of the same symme-

try species are required. Here, all but the lowest state are described by higher roots of the CI secular equations. Better zeroth-order descriptions are obtained if molecular orbitals are optimized separately for each state. However, this means that the resulting CI wavefunctions are interacting and nonorthogonal (complicating, for example, the evaluation of transition moments). The states can be made noninteracting and orthogonal by carrying out a non-orthogonal CI procedure,[77,145–148] which requires the matrix **S** of overlaps between N-electron functions and a more complex procedure for evaluating matrix elements of the Hamiltonian (the Slater-Condon rules no longer apply because they assume a single set of orthonormal orbitals). Using orbitals optimized separately for each excited state should allow one to use smaller CI expansions to achieve a given level of accuracy.[145,149]

However, optimizing excited state orbitals can be difficult because variational optimization always finds the lowest solution of a given symmetry species; this problem is generally called "variational collapse".[150] One solution is to first obtain the SCF ground state solution, and then obtain the first excited state solution by requiring it to remain orthogonal to the ground state;[151] this process could in principle be repeated for higher-lying excited states. Another solution is to optimize the orbitals by following a higher root of the MCSCF secular equations. An early application of this idea was presented by Bauschlicher and Yarkony,[85] who optimized orbitals for the $2\ ^1A_1$ state of methylene by following the second root in a two-configuration SCF procedure. A correlated treatment of this state was obtained by solving for the second root of a two-reference CISD. This same procedure, in conjunction with more highly correlated CI methods, was recently used to re-examine the $2\ ^1A_1$ state of methylene.[86] In 1987, Allen and Schaefer presented analytic gradients for this type of TCSCF-CI procedure and used them to study the $2\ ^1A_1$ state of formaldehyde and ketene.[150,152]

Unfortunately, the MCSCF optimization generally worsens the description of the ground state while it improves the description of the excited state. Frequently it happens that the energetic ordering of the two states will become swapped, in a process called "root flipping," and further optimization will yield orbitals describing the ground state.[53] One way around these difficulties is to use a single set of orbitals for all the states of a given symmetry. The improved virtual orbitals (IVO) and modified virtual orbital (MVO) methods described in the preceding section may be useful in this respect. A more typical approach is to modify the MCSCF method to yield a set of compromise orbitals; these can be obtained by the "state-averaged" procedure, which optimizes an averaged MCSCF energy obtained from averaged one- and two-electron reduced density matrices.[54] A related possibility is to use averaged natural orbitals (NOs).[153] Finally, one might simply use ground state orbitals

in conjunction with a CI method including a sufficiently complete treatment of electron correlation that the choice of orbitals becomes less important.

The most commonly used CI procedure, CISD based on the ground state configuration and using SCF orbitals, yields excitation energies which are substantially too large. One reason for this is that the ground state is correlated by all singles and double substitutions, whereas singly excited states are correlated only by singles and those doubles that involve replacement of the singly excited electron. Hence, the correlation treatment is imbalanced in favor of the ground state. This has been considered in more detail by Head-Gordon and Lee, who have analyzed the performance of CISD for excited states in the context of perturbation theory for electronic excitation energies; they find that CISD is not even correct through second order.[154] Another problem is that the SCF orbitals themselves bias results towards the ground state.

Alternatively, one might specifically design modified CI methods for excited states; the symmetry-adapted cluster (SAC) CI approach of Nakatsuji is such a method,[32,155,156] although it also contains elements from coupled-cluster theory.[6] Another alternative is the CASPT2 method of Andersson, Malmqvist, Roos, and co-workers,[157-159] which is a second-order perturbation theory based on a CASSCF reference function. Rather than employ more complex CI approaches, Pople, Head-Gordon, and others have advocated the use of configuration interaction with only singles (CIS) as a qualitative excited state theory and as a starting point for more advanced treatments.[154,160] Clearly CIS offers no improvement for the ground state (Brillouin's theorem), but higher roots represent excited states with an accuracy in the excitation energies of around 1 eV (23 kcal mol^{-1}). CIS has the unusual property of being both size extensive and variational; no other truncated CI method is size extensive. Its low computational cost and size extensivity make CIS applicable to large systems. Head-Gordon and co-workers have introduced a perturbative doubles correction for CIS which they denote CIS(D);[161] this method tends to improve excitation energies, but it does not necessarily improve geometries or other properties.[162]

The performance of simple CIS for open-shell systems such as radicals is not as good as for closed-shell systems, regardless of whether an unrestricted Hartree-Fock (UHF) or restricted open-shell Hartree-Fock (ROHF) reference is used.[163] Maurice and Head-Gordon find improved results for these systems by using a spin-pure CI wavefunction, denoted XCIS, which adds to the singles those doubly substituted determinants in which the excited electron has its spin flipped and one of the open-shell electrons is also spin flipped to conserve S_z.[164] It is interesting to note that these limited double substitutions are actually single substitutions from the point of view of spatial orbital configurations; this problem of the non-transferability of the substitution level (or

"excitation level") definition between determinants and CSFs has occasionally been mentioned in the literature.[35,165] Size extensivity is maintained in the XCIS method by uncoupling the SCF solution from the excited states. More flexible approaches which still scale favorably with system size would provide a useful alternative to the more expensive EOM-CCSD[6] and CASPT2[159] methods and are eagerly anticipated.

3 Common Features of Implementations

This section briefly discusses two elements common to all configuration interaction programs: transformation of integrals, and iterative subspace diagonalization of the Hamiltonian.

3.1 Integral Transformation

As discussed in section 2.3, the Hamiltonian matrix elements are generally written in terms of one- and two-electron integrals in the molecular orbital (MO) basis. However, these integrals are originally calculated in the atomic orbital (AO) basis, or perhaps the symmetry-adapted orbital (SO) basis. Therefore it is necessary to transform the AO or SO integrals into the MO basis, according to

$$h_{ij} = \sum_{\mu\nu} C_\mu^i C_\nu^j h_{\mu\nu}, \tag{68}$$

$$(ij|kl) = \sum_{\mu\nu\rho\sigma} C_\mu^i C_\nu^j C_\rho^k C_\sigma^l (\mu\nu|\rho\sigma), \tag{69}$$

where C_μ^i is the coefficient for the contribution of atomic orbital μ to molecular orbital i, and real orbitals have been assumed. Although the coefficients C_μ^i are generally the SCF coefficients, they might instead be the coefficients of the CI natural orbitals in the atomic orbital basis, etc.

3.1.1 One-electron Integrals

The transformation of the one-electron integrals is computationally inexpensive and easily accomplished: without point group symmetry, this transformation can be performed as two half-transformations, each of which requires a multiplication of the one-electron integral matrix by the SCF coefficient matrix, for a total of $2n^3$ multiplications. Spatial symmetry reduces this cost because the one-electron integral and SCF coefficient matrices are block diagonal according to irreducible representation (irrep), and the transformation can be carried out an irrep at a time (cf. Figure 1). Note that it would also be

Figure 1: Transformation of one-electron integrals.

Loop over irreps Γ
$h^{\Gamma}_{\mu j} = \sum_{\nu} C^{\Gamma}_{\nu j} h^{\Gamma}_{\mu\nu}$ (matrix mult)
$h^{\Gamma}_{ij} = \sum_{\mu} C^{\Gamma}_{\mu i} h^{\Gamma}_{\mu j}$ (matrix mult)
end loop over Γ

possible to utilize the permutational symmetry $h_{ij} = h_{ji}$, but this would typically reduce efficiency because the transformation could no longer be written in terms of matrix multiplications, which are very fast on vector supercomputers (e.g., CRAY C90) or pipelined workstations (e.g., IBM RS/6000).[166] Note that if orbitals are frozen in the correlated procedure (cf. section 2.4.7), $h_{\mu\nu}$ is replaced by the frozen core operator $h^{c}_{\mu\nu}$ (eq. 66).

3.1.2 Two-electron Integrals

Transforming the two-electron integrals is considerably more time-consuming. Equation (69) implies that this transformation is an n^4 process for each of n^4 integrals $(ij|kl)$ (or n^8 overall), but of course it can be carried out as four separate quarter-transformations analogous to the two half-transformations required for the one-electron integrals; this strategy requires $4n^5$ multiplications if symmetry is neglected and it constitutes a fairly demanding computational procedure if n is larger than 100 or so. Fortunately, the full transformation is not necessary for the simple CIS method because only MO integrals with two internal and two external indices are relevant. Matrix elements for most other CI wavefunctions are expressed in terms of the full set of MO integrals, but by performing some compensating work, one can avoid the full transformation for CISD,[167,168] internally-contracted MR-CISD,[101,168] and even uncontracted MR-CISD.[169] For the latter, however, Saunders and van Lenthe argue that the extra steps required to avoid the full transformation may cost more than the transformation itself unless the AO integral list exhibits considerable sparsity.[127] In the general case, the full set of integrals is required. Therefore, various methods for employing spatial and permutational symmetry have been proposed to reduce the operation count. In this context, the "permutational symmetry" refers to the eight-fold redundancy in the two-electron integrals for real orbitals.

Wilson[170] provides a very clear and helpful survey of four-index transformation methods published before 1987. More recent work has focused attention on the sparsity of quantities in the AO basis. For example, Häser, Almlöf, and Feyereisen have presented an integral-direct transformation algorithm which

Figure 2: Pre-sorting the two-electron AO integrals in TRANSQT.

Form frozen core density matrix, eq. (65)
Initialize Yoshimine structure for sorting AO two-el ints
Read two-el ints from disk; form frozen core operator, eq. (66)
    and write integrals to Yoshimine buffers
Complete Yoshimine sort, ensuring $p \geq q$, $r \geq s$, but not $pq \geq rs$
Free Yoshimine pre-sort structure
Evaluate frozen core energy, eq. (67)

employs integral pre-screening techniques and can even exploit non-abelian point-group symmetry.[171] However, our attention here is focused on integral transformation routines for highly correlated CI wavefunctions, which typically means that one can consider only small molecules for which there is less benefit in exploiting sparsity in the AO basis. Of the conventional approaches discussed by Wilson, one of the most promising is the Saunders-van Lenthe algorithm,[127] which has an operation count of $\sim 25n^5/24$ (the operation count is somewhat less if the number of transformed orbitals is less than the number of AO's). Saunders and van Lenthe present an explicit algorithm for the case of no spatial symmetry.[127] However, in our experience it is not entirely straightforward to symmetry adapt their algorithm and simultaneously maintain a high degree of vectorization. On the other hand, we find it straightforward to symmetry adapt a simpler series of quarter transformations in which some of the permutational symmetry of the integrals is ignored. This simpler code remains efficient because it calls optimized matrix multiplication subroutines. This new program, TRANSQT, developed by Daniel Crawford, Justin Fermann, and the present authors, runs faster than previous transformation programs produced by this group which take more advantage of permutational symmetries.

The algorithm consists of three major parts: a pre-sort of the two-electron SO integrals, the first half-transform, and the second half-transform (Figures 2–4). To keep track of spatial symmetry, the loops over orbitals are broken up into loops over irreps of the point group and over orbitals within those irreps. The atomic orbitals are numbered consecutively within each irrep, which allows the use of relative indices (denoted by capital letters in the figures) for numbering orbitals with a fixed irrep. Molecular orbitals are also numbered this way until they are written out at the end of the transformation, when they are renumbered according to whatever order is used by subsequent programs. As seen in the figures, all multiplications which give zero by symmetry are

avoided. The use of matrix multiplications means that the algorithm takes only partial advantage of permutational symmetry. The pre-sorted integrals P use the permutational symmetries $(pq|rs) = (qp|rs) = (pq|sr) = (qp|sr)$, but they do not allow the swapping of the first pair of indices with the last pair [e.g., $(pq|rs) = (rs|pq)$]. These same permutational symmetries are employed during multiplication except that if $rsym = ssym$, then the symmetry $(pq|rs) = (pq|sr)$ is not utilized. Similar considerations apply to the second half-transform: the half-transformed integrals J are stored similarly to P, with $(pq|kl) = (pq|lk) = (qp|kl) = (qp|lk)$.

The TRANSQT algorithm employs canonical indices[57] for pairs of orbitals, such as $pq =$`ioff`$[p] + q$. These indices are useful for computing the address of an element in a symmetric matrix which is stored by writing only the lower triangle to a linear array. The array `ioff`$[p]$ contains the address of the first element in row p, and it is assumed that $p \geq q$. If orbital numbering starts from zero, then `ioff`$[0] = 0$ and `ioff`$[p] = p +$ `ioff`$[p$-1]. The memory requirements are for two matrices (A and B) with dimension equal to the number of atomic orbitals, a matrix of SCF coefficients for each irrep, two blocks which hold all two-electron integrals $(pq|rs)$ with a fixed pair of first indices pq, and various buffers associated with Yoshimine sorting. This sorting method, first described by Yoshimine[172] in 1969, is needed to sort the integrals so that they can be read sequentially in the required order. The pre-sort is necessary because the first half-transform requires all $(pq|rs)$ for a given pq, but the integrals are not stored this way in the disk file produced by our integrals program, where they posses the full eight-fold permutational symmetry. For instance, the first half-transform will require integrals such as (11|43), but this integral is only stored as (43|11) on disk. The pre-sort adds the redundant integrals $(kl|ij) = (ij|kl)$ and places them all in the correct order for reading. The second Yoshimine sort involves the half-transformed integrals: these integrals are formed in the order $(pq|kl)$, where pq is fixed. In the second half-transform, however, the program needs to read all $(kl|pq)$ for a *fixed* kl.* Since the integrals were not written in this order, they must be sorted so they can be read this way.

3.2 Iterative Techniques for Solving Hc = *E*c

Standard numerical methods exist for diagonalizing real symmetric matrices such as the Hamiltonian **H**.† However, such methods usually require the stor-

*The convention used here is to write the fixed orbital pair first in the two-electron integral. However, one must exercise caution because the half-transformed integrals do not possess the symmetry $(pq|kl) = (kl|pq)$ since a distinction must be drawn between the AO and MO pairs.

†Once again, real orbitals have been assumed, along with a nonrelativistic Hamiltonian.

Figure 3: First half-transform in TRANSQT.

```
Initialize tmp matrices A and B and buffers Pblock and Jblock
Loop psym over irreps
    Loop p over orbitals in irrep psym
        Loop qsym over irreps w/ qsym ≤ psym
            Loop q over orbitals in irrep qsym w/ q ≤ p
                pq = ioff[p] + q
                Read (pq|rs) for all rs given pq into Pblock
                Loop rsym over irreps
                    Compute ssym from psym, qsym, rsym
                    Loop r over orbitals in irrep rsym (relative idx R)
                        Loop s over orbs in ssym w/ s ≤ r (rel idx S)
                            rs = ioff[r] + s
                            A[R][S] = Pblock[rs]
                            if rsym = ssym, A[S][R] = Pblock[rs]
                        end loop over s
                    end loop over r
                    matrix multiply:
                    loop rel idx R over orbs in rsym
                        loop rel idx L over active orbs in ssym
                            loop rel idx S over orbs in ssym
                                B[R][L] = B[R][L] + A[R][S] * C^ssym[S][L]
                    end loops over S, L, R
                    matrix multiply:
                    loop rel idx K over active orbs in rsym
                        loop rel idx L over active orbs in ssym
                            loop rel idx R over orbs in rsym
                                A[K][L] = A[K][L] + (C^rsym)^T[K][R] * B[R][L]
                    end loops over R, L, K
                    loop k over active orbitals in rsym (rel idx K)
                        loop l over active orbs in ssym (rel idx L), l ≤ k
                            kl = ioff[k] + l
                            Jblock[kl] = A[K][L]
                    end loops over l, k
                    Write Jblock to Yoshimine buffers
                end loop over rsym
            end loop over q
        end loop over qsym
    end loop over p
end loop over psym
flush, close I/O files
free Pblock
Yoshimine sort half-transformed integrals J, free Yoshimine struct
```

Figure 4: Second half-transform in TRANSQT.

Loop $ksym$ over irreps
  Loop k over active orbitals in irrep $ksym$
    Loop $lsym$ over irreps w/ $lsym \leq ksym$
      Loop l over active orbitals in irrep $lsym$ w/ $l \leq k$
      $kl = ioff[k] + l$
      zero $Jblock$
      Read all $(kl|pq)$ for given kl into $Jblock$
      Loop $psym$ over irreps
      Compute $qsym$ from $ksym$, $lsym$, and $psym$
      if $(qsym > psym)$ next $psym$
        Loop $qsym$ over irreps, with $qsym \leq psym$
          Loop p over orbs in $psym$ (rel idx P)
            Loop q over orbs in $qsym$ (rel idx Q)
              $pq = ioff[p] + q$ if $p \geq q$, else $ioff[q] + p$
              $A[P][Q] = Jblock[pq]$
            end loop over q
          end loop over p
            matrix multiply:
            Loop rel idx P over orbs in $psym$
              Loop rel idx J over active orbs in $qsym$
                Loop rel idx Q over orbs in $qsym$
                  $B[P][J] = B[P][J] + A[P][Q] * C^{qsym}[Q][J]$
            End loops over Q, J, P
            matrix multiply:
            Loop rel idx I over active orbs in $psym$
              Loop rel idx J over active orbs in $qsym$
                Loop rel idx P over orbs in $psym$
                  $A[I][J] = A[I][J] + (C^{psym})^T[I][P] * B[P][J]$
            end loops over P, J, I
            Write matrix A to buffer
          End loop over $qsym$
        End loop over $psym$
      End loop over l
    End loop over $lsym$
  End loop over k
End loop over $ksym$
free $Jblock$
flush and close I/O buffers

age of **H** in core memory (if eigenvectors are also computed, then one actually needs memory to store *two* matrices of this size). If the CI includes a mere 10,000 determinants (certainly a small CI space), storing the full matrix **H** would require 800 megabytes. As of 1997, this represents a large amount of core memory (although disk storage would not be a problem). It is little consolation that the symmetry of **H** could be used to cut this requirement approximately in half. Another very important difficulty is the time required to diagonalize matrices this size or larger. Most diagonalization routines scale as $O(n^3)$, which is certainly problematic for $n \geq 10^4$. Only for smaller matrices do the standard methods become practical.

In typical applications, only the ground electronic state or perhaps a few of the low lying excited states are of interest. Hence methods which obtain only the lowest few roots of the CI matrix are greatly preferred over methods which compute the entire spectrum. Furthermore, storage requirements are greatly reduced if **H** is not stored at all; *direct CI* methods, discussed in section 4, form products $\mathbf{Hc} = \sigma$ directly from the MO integrals.

Most techniques for solving large eigenvalue problems fall under the category of *subspace iteration* methods, which iteratively solve the eigenvalue problem in a linear vector subspace spanned by only a few vectors. Malmqvist[173] provides a concise review of the subspace iteration methods most commonly found in quantum chemistry. Here we will outline some of these methods and note recent advances.

3.2.1 Davidson's Method

Davidson's method for the iterative solution of the lowest few eigenvalues and eigenvectors of large real, symmetric matrices[108] is undoubtedly the most widely used technique for solving the CI secular equations. In this method, one applies standard diagonalization methods to a small Hamiltonian matrix formed in a subspace $\{\mathbf{b}_i\}$ of L orthonormal expansion vectors, where L increases from iteration to iteration but is typically very much smaller than the dimension of **H** (the subspace generally includes less than a dozen vectors per root). At each iteration, the Davidson algorithm estimates a correction vector for each root currently under consideration and adds it to the set $\{\mathbf{b}_i\}$ after Schmidt orthogonalization.

Davidson used perturbation theory to argue[108] that the best correction vector δ to the current iteration's guess vector **c** satisfies

$$(\mathbf{H} - \lambda\mathbf{I})\delta = -(\mathbf{H} - \lambda\mathbf{I})\mathbf{c}. \tag{70}$$

In the Davidson method, one approximates λ by the current iteration's eigenvalue, and **H** is assumed to be diagonally dominant so that δ can be approxi-

mated by

$$\delta = -(\mathbf{H}_d - \lambda\mathbf{I})^{-1}(\mathbf{H} - \lambda\mathbf{I})\mathbf{c}, \tag{71}$$

where $\mathbf{H}_d$ is the diagonal of $\mathbf{H}$ and the denominator is referred to as the *preconditioner.*

Liu showed how to extend Davidson's method to solve for several roots simultaneously,[174] leading to what is called the Simultaneous Expansion Method, the Davidson-Liu method, or the block Davidson method. The detailed Davidson-Liu algorithm, adapted from ref. 174, is presented in Figure 5.

At each iteration, the current approximations to the eigenvalues of $\mathbf{H}$ are given by the eigenvalues of the small matrix $\mathbf{G}$, which is the Hamiltonian in the subspace spanned by the expansion vectors $\{\mathbf{b}_i\}$, with matrix elements $G_{ij} = (\mathbf{b}_i, \mathbf{H}\mathbf{b}_j)$. Likewise, the current approximate eigenvectors are linear combinations of the subspace vectors with coefficients given by the eigenvectors α of $\mathbf{G}$:

$$\mathbf{c}^k = \sum_{i=1}^{L} \alpha_i^k \mathbf{b}_i. \tag{76}$$

The convergence of the k-th root can be checked by the sum of squares of the last m components of α^k in step 2 or by the norm of the residue vector $\mathbf{r}$ in step 3.

Unless very tight convergence criteria are specified, it is possible for the Davidson-Liu method to converge on the wrong eigenvector if the initial guess vectors are poor. Although this will not happen for the ground state unless a completely inappropriate guess is provided, it occasionally happens when several roots are sought. Davidson and co-workers recommend initial loose convergence of more roots than are actually needed, and then tighter convergence on the desired roots.[175] Possible choices for the initial vectors include unit vectors (chosen according to the diagonal elements of $\mathbf{H}$ with the largest magnitudes) or eigenvectors of some small block of $\mathbf{H}$.

Equations (73) and (75) show that after a matrix-vector product $\sigma_i = \mathbf{H}\mathbf{c}_i$ is computed it is needed again in subsequent iterations. Since the construction of the vectors $\{\sigma_i\}$ is the most time consuming step in the iterative diagonalization of $\mathbf{H}$, they are stored on disk along with the expansion vectors $\{\mathbf{b}_i\}$. The original Davidson method converges one root at a time and requires storage for two (segments of) vectors at once in core memory. If more core memory is available, then the Davidson-Liu method can reduce computational and I/O requirements. For example, one pass through the subspace expansion vectors is sufficient to construct several correction vectors δ simultaneously. Likewise, a single construction of the matrix elements of $\mathbf{H}$, several σ vectors can be formed simultaneously.

Note that the preconditioner in eq. (74) requires the diagonal elements of

Figure 5: The Davidson-Liu Iterative Method for the Lowest Few Eigenvectors and Eigenvalues of Real, Symmetric Matrices (Ref. 174)

1. Select a set of L orthonormal guess vectors, at least one for each root desired, and place in the set $\{\mathbf{b}_i\}$.

2. Use a standard diagonalization method to solve the $L \times L$ eigenvalue problem
$$\mathbf{G}\alpha^k = \lambda^k \alpha^k, k = 1, 2, \ldots, M \tag{72}$$
where
$$G_{ij} = (\mathbf{b}_i, \mathbf{H}\mathbf{b}_j) = (\mathbf{b}_i, \sigma_j), 1 \leq i, j \leq L \tag{73}$$
and M is the number of roots of interest.

3. Form the correction vectors $\{\delta^k\}, k = 1, 2, \ldots, M$, defined as
$$\delta_I^k = (\lambda^k - H_{II})^{-1} r_I^k, I = 1, 2, \ldots, N \tag{74}$$
where
$$\mathbf{r}^k = \sum_{i=1}^{L} \alpha_i^k (\mathbf{H} - \lambda^k)\mathbf{b}_i \tag{75}$$
and N is the number of determinants or CSFs.

4. Normalize $\{\delta^k\}$.

5. Schmidt orthonormalize δ^1 against the set $\{\mathbf{b}_i\}$ and append the result to $\{\mathbf{b}_i\}$. Repeat this process for each of the other $M-1$ correction vectors, neglecting those whose Schmidt orthonormalized norm is less than some threshold $T \sim 10^{-3}$. This results in the addition of m new $\mathbf{b}$ vectors, with $1 \leq m \leq M$.

6. Increase L by m and return to step 2.

the Hamiltonian. These can be precomputed and stored on disk, or they can be computed on-the-fly. Alternatively, they can be approximated in some cases using orbital energies. In a determinantal basis, Davidson's preconditioner can actually cause the CI vector to break spin symmetry. Indeed, Knowles and Handy noted this difficulty in their pioneering 1984 paper on determinant based configuration interaction.[109] They found that this problem can be avoided if the diagonal elements of the Hamiltonian H_{II} are replaced by an average $\tilde{H}_{II}$ over all determinants which have the same spatial orbital configuration as determinant I but differ in the distribution of spins.

If several roots are sought, or convergence takes many iterations, then the number of vectors stored on disk can become large, leading to I/O delays during the construction of $\mathbf{G}$ or $\mathbf{r}$. Furthermore, disk storage can become a problem if the vectors themselves are very large. One solution is to apply compression algorithms to the subspace vectors and σ vectors.[176,177] Additionally, the Davidson-Liu method can be restarted with only M expansion vectors by using the current approximation to each eigenvector, eq. (76), as the new starting guess vectors. Clearly this procedure hinders the rate of convergence because information is lost after the vector subspace is collapsed: subsequent diagonalizations have less variational freedom because of the reduced dimension of the subspace. However, in 1990 van Lenthe and Pulay presented the remarkable conclusion[178] that when only a single root is sought, collapsing the subspace does not substantially degrade the rate of convergence if the subspace is collapsed down to two vectors instead of just one. This procedure, which may be justified by the theory of conjugate gradients, was later generalized to multiple roots by Murray, Racine, and Davidson.[175] The collapsed vector subspace contains the current guess vector for each root, as before, and also the guess vectors from the previous iteration (after they have been Schmidt orthogonalized against the other vectors in the collapsed subspace).

Other work has focused on improving the correction vector. As noted by Olsen,[83] Saad,[179] Sleijpen and van der Vorst,[180] and others,[181,182] Davidson's equation (70) seems to imply that the optimal update vector δ is just the negative of the current guess CI vector $\mathbf{c}$. Clearly, this would not allow for the expansion of the vector subspace. Sleijpen *et al.*[180] have pointed out that Davidson assumed that δ is orthogonal to $\mathbf{c}$ in deriving eq. (70). However, Davidson's method only enforces this orthogonality after δ has already been determined. Hence, a more effective preconditioner may result from explicitly enforcing this orthogonality while δ is being constructed, and several authors have recently proposed such preconditioners.[83,181,182]

Another improvement suggested by these authors and others[54,56,173] is to lift the assumption of strict diagonal dominance of the Hamiltonian in the preconditioner. One selects a subspace of the most important N-electron functions

in the CI space and, for the purposes of the preconditioner, approximates the Hamiltonian as

$$\mathbf{H} \approx \begin{bmatrix} \mathbf{H}' & 0 \\ 0 & \mathbf{\Lambda}'' \end{bmatrix}, \tag{77}$$

where primes denote the small selected subspace and double primes denote the complement subspace. Although relatively few (up to several hundred) determinants might be included in the selected space, it is important to take complete sets of determinants which are capable of forming spin eigenfunctions so that the spin symmetry of the CI vector can be maintained during the iterative procedure. Note that the Hamiltonian is assumed diagonal in the complement subspace and coupling is ignored. This leads to two equations for the correction vector:

$$(\delta^k)' = -(\mathbf{H}' - \lambda^k \mathbf{I}')^{-1} (\mathbf{r}^k)' \tag{78}$$

$$(\delta^k)'' = -(\mathbf{\Lambda}'' - \lambda^k \mathbf{I}'')^{-1} (\mathbf{r}^k)''. \tag{79}$$

The second equation of course becomes

$$(\delta_I^k)'' = \frac{(r_I^k)''}{\lambda_k - H_{II}}, \tag{80}$$

and is analogous to eq. (74). The first equation can be written in terms of the eigenvalues μ^l and eigenvectors $\mathbf{u}^l$ of the small matrix $\mathbf{H}'$:

$$(\delta_I^k)' = \sum_l \frac{u_I^l (\mathbf{u}^l \cdot (\mathbf{r}^k)')}{\lambda^k - \mu^l} \tag{81}$$

3.2.2 Olsen's Method

Realizing the difficulties of storing several $\mathbf{b}$ and σ vectors for very large CI spaces, Olsen proposed that each correction vector be added directly to the current CI vector, and that the resulting (renormalized) vector be used as the next iteration's guess vector. Of course for this scheme to work well, the correction vector must be as good as possible. Olsen therefore introduced an improved method for generating the correction vector, using some of the ideas just discussed above. If the current (normalized) CI vector is denoted $\mathbf{c}$, then its energy is

$$E = (\mathbf{c}, \mathbf{Hc}). \tag{82}$$

The Hamiltonian is then divided into two terms,

$$\mathbf{H} = \mathbf{H}^{(0)} + \mathbf{H}^{(1)}, \tag{83}$$

and the CI eigenvalue equation can be written

$$(\mathbf{H}^{(0)} + \mathbf{H}^{(1)})(\mathbf{c} + \delta\mathbf{c}) = (E + \delta E)(\mathbf{c} + \delta\mathbf{c}), \tag{84}$$

where $\delta\mathbf{c}$ and δE are the corrections to the current CI vector and energy. If $\delta\mathbf{c}$ is required to be orthogonal to $\mathbf{c}$ then (neglecting quadratic terms)

$$\delta\mathbf{c} = -(\mathbf{H}^{(0)} - E)^{-1}\left[(\mathbf{H} - E)\mathbf{c} - \mathbf{c}\delta E\right], \tag{85}$$

where

$$\delta E = \frac{(\mathbf{c}, (\mathbf{H}^{(0)} - E)^{-1}(\mathbf{H} - E)\mathbf{c})}{(\mathbf{c}, (\mathbf{H}^{(0)} - E)^{-1}\mathbf{c})} \tag{86}$$

The correction vector $\delta\mathbf{c}$ is superior to that used in the standard Davidson method since it remains rigorously orthogonal to $\mathbf{c}$ and therefore retains the ability to introduce new character into the CI vector even near convergence.[83] This correction vector was also derived by Bofill and Anglada from other considerations.[181] In 1990, Olsen, Jørgensen, and Simons used this method to perform three iterations of the first one-billion determinant CI calculation.[83] The zeroth-order part of the Hamiltonian $\mathbf{H}^{(0)}$ was defined as a 400×400 block of determinants formed from the lowest diagonal elements of $\mathbf{H}$, and as the diagonal of $\mathbf{H}$ outside this block. This procedure requires the storage of four vectors on disk (three if the diagonal elements of $\mathbf{H}$ are computed as needed).

Although the Olsen method can be very helpful when disk space is limited, its convergence characteristics are not always very good. Indeed, as first pointed out by Mitrushenkov,[183] the Olsen method does not guarantee that the energy decreases every iteration. However, it is of course possible to use Olsen's preconditioner in conjunction with iterative methods which keep more than one CI vector and σ vector. Mitrushenkov[183] advocates diagonalizing the Hamiltonian in the space of the current and previous CI vectors:

$$H_{ii} = (\mathbf{c}^{(i)}, \sigma^{(i)}) \tag{87}$$

$$H_{i,i-1} = H_{i-1,i} = (\sigma^{(i)}, \mathbf{c}^{(i-1)}). \tag{88}$$

Explicitly, the nonorthogonal pseudo-eigenvalue equation is

$$\begin{pmatrix} H_{i-1,i-1} & H_{i-1,i} \\ H_{i,i-1} & H_{i,i} \end{pmatrix} \begin{pmatrix} \alpha_{i-1} \\ \alpha_i \end{pmatrix} = E_i \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix} \begin{pmatrix} \alpha_{i-1} \\ \alpha_i \end{pmatrix}, \tag{89}$$

where s is the overlap between the current and previous CI vectors, $(\mathbf{c}^{(i)}, \mathbf{c}^{(i-1)})$. At each iteration, the current CI vector is recomputed as $\mathbf{c}^{(i)} = \alpha_{i-1}\mathbf{c}^{(i-1)} + \alpha_i\mathbf{c}^{(i)}$, and $\sigma^{(i)} = \alpha_{i-1}\sigma^{(i-1)} + \alpha_i\sigma^{(i)}$. H_{ii} is set to E_i in eq. (89), and then the new vector $\mathbf{c}^{(i+1)}$ is computed using Olsen's method. This process is repeated until convergence is reached. In our experience Mitrushenkov's method

improves convergence substantially in the first few iterations compared to the single-vector version of Olsen's method. Unfortunately, as CI vector approaches convergence, eq. (89) becomes ill-conditioned because $H_{i-1,i}$ approaches H_{ii} and s approaches unity. This difficulty can be avoided by reverting to the single-vector Olsen method near convergence.

4 Determinant-Based Algorithms for Highly Correlated CI

This section describes several determinant-based CI algorithms. The alpha and beta string formalism of Handy[44] is introduced, and the equations for $\sigma = \mathbf{Hc}$ within this formalism are derived for the full CI and RAS CI cases.[46] Practical considerations for implementation are also discussed.

4.1 Slater Determinants, CSFs, and Direct CI

Slater determinants are eigenfunctions of $\hat{S}_z$; therefore, the CI space includes only those determinants with a given value of M_s unless a spin-dependent Hamiltonian is used.[184] However, Slater determinants are not eigenfunctions of $\hat{S}^2$ as are configuration state functions (CSFs), and dimension of the CI space in Slater determinants is typically about 2-4 times larger than in CSFs (for more about CSFs, see the books by Pauncz[185,186]). Thus it would seem that CSFs are preferable to Slater determinants for use as CI expansion functions. However, determinants offer certain advantages in the context of "direct CI" methods that can outweigh the disadvantage of a larger CI space.

In computational quantum chemistry the term "direct" has come to mean that certain quantities, which are too large to hold in core memory, are computed on-the-fly instead of being stored on disk and read as needed. A *direct SCF* implies that the two-electron integrals are computed on-the-fly. For a *direct CI*,[88] the integrals are still held on disk, but the Hamiltonian matrix itself is not explicitly constructed or stored. Instead, the vector $\sigma = \mathbf{H}c$, which is required in iterative subspace methods for diagonalizing the Hamiltonian (cf. section 3.2), is computed directly from the one- and two-electron integrals and the CI vector; the construction of σ is the time-consuming step in the direct CI method. The coefficients for multiplying the CI vector by the integrals have already been introduced (cf. section 2.3.2) as the one- and two-electron coupling coefficients; these may be written to disk in a file traditionally called the "formula tape." Unfortunately, this procedure is still unsuitable for a direct CI, since the coupling coefficients will require as much storage space as the Hamiltonian matrix itself, leading to long I/O delays. Hence, in a direct

CI, the coupling coefficients should be calculated as needed (or they should be built into the program). A next step would be to eliminate storage of the CI vector itself; efforts along these lines have been described as *superdirect CI*.[187,188] Alternatively, Carter and co-workers have considered pseudospectral approaches which eliminate the need to construct two-electron integrals as separate intermediates.[189–192]

The direct CI method was first introduced in 1972 by Roos for the case of CISD from a closed-shell reference function.[193] However, generalization to more complex CI spaces, such as MR-CISD, proved exceedingly difficult due to the large number of special cases. The next breakthrough did not occur until Shavitt[41,42] cast the work of Paldus on the unitary group approach (UGA)[40,78] into a graphical formalism which represents CSFs in the Gelfand-Tsetlin canonical basis as walks on a directed graph. Not only did this graphical representation make the UGA more accessible to chemists, but it also provided a convenient formalism for carrying out computations. Any pair of walks (CSFs) forms a loop, and matrix elements are evaluated based on the shapes of these loops, with only certain loop types giving nonzero matrix elements.[41,42] For example, one-electron coupling coefficients are expressed as

$$\gamma_{ij}^{IJ} = \prod_{k=i}^{j} W(T_k, b_k), \tag{90}$$

where T_k identifies the shape of the loop formed by walks I and J at level k on the Shavitt graph, and b_k is the "b-value" of the vertex on walk J at level k (see ref. 42 for more details). The coupling coefficient vanishes unless walks I and J coincide everywhere below level $i-1$ and above level j on the graph (assuming $i < j$). This graphical unitary group approach (GUGA) was developed with a philosophy similar to that of the direct CI. The idea was to use each coupling coefficient, specified by a loop on the Shavitt graph, for a whole series of Hamiltonian matrix elements differing in their common upper and lower walks. The first multireference CI method based on the ideas of Paldus and Shavitt was developed by Brooks and Schaefer.[43] Particularly notable was their computation on the $^1B_{1u}$ state of ethylene involving all single and double excitations relative to three open-shell singlet reference configurations.[43] However, a detailed analysis of the "loop-driven" GUGA CI program of Brooks and Schaefer indicates that, in practice, few loops contribute to very many matrix elements, and it remains more efficient to write the coupling coefficients to the formula tape rather than to recompute them as needed.

Nevertheless, the graphical approach afforded new insight into the structure of the Hamiltonian. In particular, for CI spaces which allow only one or two electrons in the external space, the graphical representation of the external space becomes very regular, and the external portion of a loop can only have

a few possible shapes.[194] Siegbahn made the crucial observation that the one- and two-electron coupling coefficients can be factored into contributions from the internal orbitals and from the external orbitals,

$$\gamma_{ij}^{IJ} = {}^{int}\gamma_{ij}^{IJ} \times {}^{ext}\gamma_{ij}^{IJ} \tag{91}$$

$$\Gamma_{ijkl}^{IJ} = {}^{int}\Gamma_{ijkl}^{IJ} \times {}^{ext}\Gamma_{ijkl}^{IJ}, \tag{92}$$

and that the calculation can be "direct" in the external space when the external factors are very simple.

In 1979, Siegbahn showed[195] that for the case of single replacements from a reference wavefunction which is a full CI in the active space (i.e., first-order CI[13]), the external factors for the coupling coefficients are all simply +1; hence, only the internal space coupling coefficients must be precomputed and stored on the formula tape. This results in a substantial savings because the number of internal coupling coefficients will be much smaller than the total number of coupling coefficients. In 1980, Siegbahn extended these ideas to the general case of all single and double substitutions for an arbitrary set of references (i.e., MR-CISD).[194] Once again, the external coupling coefficient factors are simple (± 1, $\pm\sqrt{2}$, and 2) and can be dealt with in a direct fashion. The shape-driven GUGA program of Saxe, Fox, Schaefer, and Handy[91] is based in part on Siegbahn's approach, as is the program of Saunders and van Lenthe[127] and the COLUMBUS program of Shavitt, Lischka, Shepard, and co-workers.[196–198]

Unfortunately, these simplifications are not directly applicable[91] when more than two electrons are allowed into the external space (e.g., CISDT, CISDTQ, and full CI). Furthermore, even for MR-CISD wavefunctions, large active spaces can lead to a large number of internal coupling coefficients, which can become difficult to deal with.[197] The next advance was once again due to Siegbahn, who suggested a factorization of the two-electron coupling coefficients by inserting the resolution of the identity:

$$\Gamma_{ijkl}^{IJ} = \sum_K \gamma_{ij}^{IK}\gamma_{kl}^{KJ} - \gamma_{il}^{IJ}\delta_{jk}. \tag{93}$$

Although the resolution of the identity requires an infinite sum in principle, in this case only a finite number of states $|\Phi_K\rangle$ are relevant. For fixed i, j, k, l, I, and J, the product term will vanish unless $|\Phi_K\rangle$ is obtained from $|\Phi_J\rangle$ by a single substitution from orbital l to orbital k and from $|\Phi_I\rangle$ by a single substitution from orbital i to orbital j. For configuration state functions, this completely specifies the orbital configuration of $|\Phi_K\rangle$, and only a few spin couplings must be summed over. This approach led Knowles and Handy[109] to present a vectorized full CI algorithm based on Slater determinants rather than CSFs. For Slater determinants, the one-electron coupling coefficients

appearing in (93) and elsewhere are very easy to calculate on-the-fly, allowing a fully direct CI procedure. For Slater determinants $|I\rangle$ and $|J\rangle$, $\gamma_{ij}^{IJ} = \langle I|\hat{E}_{ij}|J\rangle$ is 0 unless determinant $|J\rangle$ becomes determinant $|I\rangle$ (within a sign) when an alpha or beta electron is moved from orbital j to orbital i, in which case γ_{ij}^{IJ} becomes ± 1. For the special case $i = j, I = J$, γ_{ii}^{II} counts the number of electrons in orbital i for determinant I, yielding 0, 1, or 2.

Because of this simplicity and the ability to carry out computations in a fully direct fashion, many of the full and restricted CI algorithms developed over the last ten years have employed Slater determinants, and in this section we will focus our attention on these determinant-based methods. However, we should note that further progress has also been made in CSF-based approaches. Much of the recent work in direct CI methods has been based on the symmetric group approach (SGA)[199–203] rather than the related unitary group approach (UGA).[40–42,78] Given the factorization (93), the problem of formulating a fully direct CI procedure can be turned into the problem of determining the one-electron coupling coefficients on-the-fly in the desired order. Knowles and Werner presented a way of doing this in 1988.[102] They use the identity

$$\hat{E}_{ij} = \hat{E}_{ia}\hat{E}_{aj}, \tag{94}$$

which holds for any orbital ϕ_a which is always unoccupied. This hypothetical orbital, referred to as a "ghost" orbital, does not actually appear in any of the integrals. The one-electron coupling coefficient becomes

$$\gamma_{ij}^{IJ} = \sum_{K_s} \langle I|\hat{E}_{ia}|K_s\rangle\langle K_s|\hat{E}_{aj}|J\rangle, \tag{95}$$

where the sum is over all spin couplings of the uniquely-specified orbital configuration K. By fixing one of the two orbital indices, it becomes feasible to store the intermediates needed to evaluate the one-electron coupling coefficients efficiently in the desired order. This ghost-orbital technique was used by Werner and Knowles in their implementation of internally-contracted multireference CI.[101] That method requires third- and fourth-order reduced density matrices, which can be evaluated by approaches analogous to (93). Another possibility along the lines of Siegbahn's internal/external factorization has been suggested by Malmqvist, Rendell, and Roos in their implementation of the RAS SCF method.[56] They modify the GUGA method to split all walks into upper and lower portions and calculate coupling coefficients as products of upper and lower factors. Although the upper factors are not necessarily very simple for a RAS case, the storage requirements are substantially reduced in this approach.

Finally, we note that many other important advances have been made in CSF-based approaches, even outside the context of direct CI. Much of this effort in recent years has focused on extending the unitary and symmetric group

approaches to the spin-dependent Hamiltonians needed to account for relativistic effects.[204-210] Examples of other work include specialized unitary group approaches for MRCI wavefunctions based on CAS references[211] and application of the unitary group approach to CI calculations on atoms using Hylleraas coordinates[65] and to spin-adapted open-shell coupled-cluster theory.[212] However, we now turn our attention to determinant-based formulations of direct CI.

4.2 Alpha and Beta Strings

A 1980 paper by Handy[44] represented a major advance in determinant-based CI, even though the paper was more concerned with how integrals and CI coefficients are stored than with the computational advantages of determinants over CSFs. Handy realized that if determinants are used as N-electron basis functions, and particularly if these determinants are expressed as "alpha strings" and "beta strings," then the vector σ can be evaluated very efficiently.

Although Handy was the first to use alpha and beta strings, we will employ the subsequent notation of Olsen *et al.*[46] An alpha string is defined as an ordered product of creation operators for spin orbitals with alpha spin. If I_α contains a list $\{i, j, \ldots k\}$ of the N_α occupied spin orbitals with alpha spin in determinant $|I\rangle$, then the alpha string $\alpha(I_\alpha)$ is $a^\dagger_{i\alpha} a^\dagger_{j\alpha} \ldots a^\dagger_{k\alpha}$. A beta string is defined similarly. Thus a Slater determinant $|I\rangle$ in terms of alpha and beta strings is

$$|I\rangle = |\alpha(I_\alpha)\beta(I_\beta)\rangle = \alpha(I_\alpha)\beta(I_\beta)|\rangle. \tag{96}$$

For example, consider the Slater determinant $|I\rangle = |\phi_{1\alpha}\phi_{2\alpha}\phi_{3\alpha}\phi_{1\beta}\phi_{2\beta}\phi_{4\beta}\rangle$. Then the alpha string $\alpha(I_\alpha)$ is given by

$$\alpha(I_\alpha) = a^\dagger_{1\alpha} a^\dagger_{2\alpha} a^\dagger_{3\alpha}, \tag{97}$$

and the beta string is given by

$$\beta(I_\beta) = a^\dagger_{1\beta} a^\dagger_{2\beta} a^\dagger_{4\beta}. \tag{98}$$

Note that the order of the creation operators matters; if we swap the order of two creation operators within the alpha string (or within the beta string), then we introduce a sign change due to the anticommutation relation of creation operators. Also, applying the alpha string to the vacuum first, rather than the beta string, may introduce a minus sign, depending on the number of alpha and beta electrons. Typically, the beta string will be placed to the right of the alpha string in equations like (96). Further, within each string, orbitals are listed in strictly increasing order.

Since a determinant is now specified by an ordered pair of indices representing its alpha and beta components, the Ith element of the CI vector becomes $c(I_\alpha, I_\beta)$. Note that this vector can also be considered as a matrix with coordinates I_α and I_β. Both vector and matrix addressing schemes are computationally useful. The σ vector,

$$\sigma_I = \sum_J H_{IJ} c_J, \tag{99}$$

can be written in the new notation as

$$\sigma(I_\alpha, I_\beta) = \sum_{J_\alpha, J_\beta} \langle \beta(J_\beta)\alpha(J_\alpha)|\hat{H}|\alpha(I_\alpha)\beta(I_\beta)\rangle c(J_\alpha, J_\beta), \tag{100}$$

where we have used $H_{IJ} = H_{JI}^*$ and assumed that $\mathbf{H}$ is a real matrix to obtain the form most commonly seen in the literature when this notation is used. Handy realized the following advantages to alpha and beta strings:

1. Direct CI methods often require an index vector which points to a list of all allowed excitations from a given N-electron basis function. Using alpha and beta strings, the index vector need not be the length of the CI vector—its size is dictated by the number of alpha or beta strings, which (for a full CI) is approximately the square root of the number of determinants. This results from the fact that in determinant-based CI, electrons in alpha spin-orbitals can be excited only to other alpha spin-orbitals, and electrons in beta spin-orbitals can be excited only to other beta spin-orbitals (because of the restriction to a single value of M_s).

2. To form $\sigma(I_\alpha, I_\beta)$ in equation (100), all functions $|\alpha(J_\alpha)\beta(J_\beta)\rangle$ which have non-zero matrix elements with $|\alpha(I_\alpha)\beta(I_\beta)\rangle$ are generated, one at a time, with the appropriate integral being looked up and multiplied by the appropriate CI coefficient. No time is wasted considering determinants which are noninteracting, and the coefficients of the integrals are easy to evaluate.

3. Efficiency is increased by realizing that all integrals which enter the expression $\langle \alpha(I_\alpha)\beta(I_\beta)|\hat{H}|\alpha(J_\alpha)\beta(I_\beta)\rangle$ (equation 100), where $\alpha(J_\alpha)$ differs from $\alpha(I_\alpha)$ by two orbitals, are independent of $\beta(I_\beta)$.

This approach allowed several benchmark full CI computations, including the first CI procedure (1981) to include more than one million determinants.[38,213]

4.3 The Vectorized Full CI Algorithm of Knowles and Handy

In 1984, Knowles and Handy introduced a new direct CI algorithm for full CI wavefunctions.[109,214] As Siegbahn had pointed out,[45] the efficiency of direct CI algorithms is increased if, for a given $|K\rangle$, all one-electron coupling coefficients γ_{kl}^{KJ} are available together. Examination of Siegbahn's expression for σ elucidates this observation:[45]

$$\sigma_I = \sum_J H_{IJ} c_J \tag{101}$$

$$= \sum_{ij} \sum_J \gamma_{ij}^{IJ} h_{ij} c_J + \frac{1}{2} \sum_{ijkl} (ij|kl) \sum_J \Gamma_{ijkl}^{IJ} c_J \tag{102}$$

$$= \sum_{ij} \{h_{ij} - \frac{1}{2} \sum_l (il|lj)\} \sum_J \gamma_{ij}^{IJ} c_J$$

$$+ \frac{1}{2} \sum_{ijkl} (ij|kl) \sum_K \gamma_{ij}^{IK} \sum_J \gamma_{kl}^{KJ} c_J, \tag{103}$$

where the resolution of the identity has been used to turn the two-electron coupling coefficients into products of one-electron coupling coefficients (eq. 93). Notice that part of the two-electron contribution has been folded into the one-electron term. The remaining two-electron term is the time consuming part in the evaluation of σ_I, and it is most efficiently written as

$$\sigma_I^{(2)} = \frac{1}{2} \sum_K \sum_{ij} \gamma_{ij}^{IK} \sum_{kl} (ij|kl) \sum_J \gamma_{kl}^{KJ} c_J. \tag{104}$$

Thus this part of σ_I can be evaluated by the following set of operations:

$$D_{kl}^K = \sum_J \gamma_{kl}^{KJ} c_J \tag{105}$$

$$E_{ij}^K = \sum_{kl} (ij|kl) D_{kl}^K \tag{106}$$

$$\sigma_I^{(2)} = \frac{1}{2} \sum_K \sum_{ij} \gamma_{ij}^{IK} E_{ij}^K. \tag{107}$$

The one-electron coupling coefficients would ordinarily be stored on disk, making the evaluation of the D and σ quantities I/O intensive and thus inefficient. However, Knowles and Handy noted that in a basis of determinants the one-electron coupling coefficients can be evaluated on-the-fly (direct CI) even in the general case. A given determinant $|K\rangle$ can interact with at most two

other determinants $|I\rangle$, and their contributions can be separated by rewriting the shift operator in $\gamma_{ij}^{IK} = \langle I|\hat{E}_{ij}|K\rangle$ as

$$\hat{E}_{ij} = \hat{E}_{ij}^{\alpha} + \hat{E}_{ij}^{\beta}, \tag{108}$$

where $\hat{E}_{ij}^{\alpha}$ replaces an α-spin electron in orbital j with an α-spin electron in orbital i (cf. section 2.3.2). Note that this approach requires a sum over a complete set of intermediate determinants $|K\rangle$ (or at least all determinants which interact with the allowed determinants through $\hat{E}_{ij}$), including those determinants with the wrong spatial symmetry. This means that the Knowles and Handy algorithm would be considerably less efficient for restricted CI.

Given equation (108), it is possible to write the one-electron coupling coefficients in terms of alpha and beta strings:

$$\begin{aligned} \gamma_{ij}^{IJ} &= \langle\alpha(I_\alpha)\beta(I_\beta)|\hat{E}_{ij}^{\alpha}|\alpha(J_\alpha)\beta(J_\beta)\rangle\delta(I_\beta, J_\beta) \\ &+ \langle\alpha(I_\alpha)\beta(I_\beta)|\hat{E}_{ij}^{\beta}|\alpha(J_\alpha)\beta(J_\beta)\rangle\delta(I_\alpha, J_\alpha), \end{aligned} \tag{109}$$

where $|\alpha(I_\alpha)\rangle$ is related to $|\alpha(J_\alpha)\rangle$ by a single excitation (and likewise for $|\beta(I_\beta)\rangle$ and $|\beta(J_\beta)\rangle$). Thus the one-electron coupling coefficients are generated from lists of strings related by single excitations. For each alpha (beta) string, one stores a list of all allowed single replacements to other alpha (beta) strings; for closed-shell systems, the two lists will be identical and only one must be stored. Each list contains the address of the new string, the orbital index ij, and a phase factor, denoted by $\mathrm{sgn}(ij)$, which is ± 1. The sign can be determined as $(-1)^p$, where p is the number of transpositions of creation operators needed to bring an excited string to its canonical form. String addresses were computed by table lookups using a canonical addressing scheme explained in section 4.9.2. One can take advantage of the permutational symmetries $(ij|kl) = (ji|kl) = (ij|lk) = (ji|lk)$ of the two-electron integrals by requiring $i \geq j$, $k \geq l$. This entails replacing γ_{kl}^{KJ} in (105) by $(\gamma_{kl}^{KJ} + \gamma_{lk}^{KJ})$, an analogous change for (107), and a modification of the integrals to avoid double counting when $i = j$ or $k = l$.

The algorithm of Knowles and Handy is described as "vectorized" because each of the three major operations (105)-(107) may be written as an operation performed on an entire vector at once. This is very beneficial for vector supercomputers, which actually perform such operations a vector at a time and give substantial increases in speed. To illustrate, consider Fig. 6, which shows the Knowles-Handy algorithm for the formation of D, eq. (105). Due to memory limitations, operations are performed for a block of strings at a time. In the first half of Fig. 6, the operations in the innermost loop are identical but independent of each other for different K_β. In the second half of the algorithm, the same applies to K_α; hence, this operation can be performed for a

Figure 6: Knowles-Handy Vectorized Formation of D (Refs. 109,214).

loop alpha strings K_α in block
  loop over excitations $|\alpha(J_\alpha)\rangle = \mathrm{sgn}(ij)\hat{E}^\alpha_{ij}|\alpha(K_\alpha)\rangle$
    loop over K_β in block
      $|K\rangle = |\alpha(K_\alpha)\beta(K_\beta)\rangle, |J\rangle = |\alpha(J_\alpha)\beta(K_\beta)\rangle$
      $D(K_\alpha, K_\beta, ij) = D(K_\alpha, K_\beta, ij) + \mathrm{sgn}(ij)c(J_\alpha, K_\beta)$

loop alpha strings K_β in block
  loop over excitations $|\beta(J_\beta)\rangle = \mathrm{sgn}(ij)\hat{E}^\beta_{ij}|\beta(K_\beta)\rangle$
    loop over K_α in block
      $|K\rangle = |\alpha(K_\alpha)\beta(K_\beta)\rangle, |J\rangle = |\alpha(K_\alpha)\beta(J_\beta)\rangle$
      $D(K_\alpha, K_\beta, ij) = D(K_\alpha, K_\beta, ij) + \mathrm{sgn}(ij)c(K_\alpha, J_\beta)$

whole range of K_β (K_α) values simultaneously with a vector processor. These same considerations apply to the analogous eq. (107). The remaining (and most time consuming) step, (106), can be performed as a matrix multiplication when one uses compound indices ij and kl, and of course this is also a vectorized operation.

These concerns about vectorization remain relevant even though quantum chemists now perform a substantial fraction of their computations on workstation machines which lack vector processors. Nevertheless, workstations (and now even personal computers, or PCs) feature *pipelined* processors. Pipelines allow machine instructions to overlap to some extent, giving the processor a limited ability to perform several tasks at once.[166] The superscalar IBM RS/6000 POWER2 workstation processor has two floating point pipelines, each of which can hold up to twelve instructions. If the processor can keep a steady stream of independent instructions coming down the pipeline, then overall performance will be increased substantially. However, if one instruction depends on results from another, then the pipeline can become stalled and performance is degraded. Since vectorizable code implies many similar but independent operations, as a general rule, vectorizable code becomes good pipelined code.

The Knowles-Handy approach was expected to be very efficient on vector supercomputers, and indeed it enabled many important full CI benchmark calculations.[15] Nevertheless, one can see that this algorithm does more work than is strictly necessary. Equation (106) demonstrates that the operation count for the time-consuming step is approximately $\frac{1}{4}\tilde{N}_{det}n^4$, where n is the number of orbitals and $\tilde{N}_{det}$ is the number of interacting intermediate states,

which may be larger than the number of determinants in the full CI space because the intermediate states are not subject to spatial symmetry restrictions. The factor of $\frac{1}{4}$ arises from the permutational symmetries of the integrals. As discussed previously in section 2.4.5, the computational cost of a full CI procedure should actually scale as $\mathcal{O}(N_{det}N^2n^2)$. Thus, the Knowles-Handy algorithm replaces N_{det} with the larger $\tilde{N}_{det}$, and N^2 with the larger n^2. Some of the extra work is due to the fact that the intermediate matrix D can contain a substantial number of zeros. For example, D^K_{kl} will be zero when K has orbital k unoccupied or orbital l doubly occupied (D becomes less sparse under the condition $k \geq l$). Even though matrix multiplications are ideal for vector computers, Olsen and co-workers[46] realized that abandoning the matrix formulation (105)-(107) might still lead to a faster algorithm due to the substantially reduced operation count.

4.4 Olsen's String-Based Full CI Algorithm

In order to avoid the unnecessarily large operation counts in the full CI algorithm of Knowles and Handy, Olsen *et al.* abandoned the explicit use of a complete set of intermediate states and returned to some of Handy's original (1980) ideas[44] concerning string-driven full CI approaches.

4.4.1 Full CI σ Equations

We begin by describing Olsen's expressions for the σ vector. In second quantized-form (cf. section 2.3.2), $\hat{H}$ becomes

$$\hat{H} = \sum_{kl}^{n} h_{kl}\hat{E}_{kl} + \frac{1}{2}\sum_{ijkl}^{n}(ij|kl)\left(\hat{E}_{ij}\hat{E}_{kl} - \delta_{jk}\hat{E}_{il}\right). \tag{110}$$

Inserting this expression into that for σ, eq. (100), yields

$$\begin{aligned}\sigma(I_\alpha, I_\beta) &= \sum_{J_\alpha, J_\beta} \langle\beta(J_\beta)\alpha(J_\alpha)|\sum_{kl}^{n} h_{kl}\hat{E}_{kl} \\ &+ \frac{1}{2}\sum_{ijkl}^{n}(ij|kl)\left(\hat{E}_{ij}\hat{E}_{kl} - \delta_{jk}\hat{E}_{il}\right)|\alpha(I_\alpha)\beta(I_\beta)\rangle c(J_\alpha, J_\beta)\end{aligned} \tag{111}$$

Now expanding the shift operators into their two spin components, $\hat{E}_{kl} = \hat{E}^\alpha_{kl} + \hat{E}^\beta_{kl}$, we write σ as a sum of three terms:[46]

$$\sigma(I_\alpha, I_\beta) = \sigma_1(I_\alpha, I_\beta) + \sigma_2(I_\alpha, I_\beta) + \sigma_3(I_\alpha, I_\beta), \tag{112}$$

where

$$\sigma_1(I_\alpha, I_\beta) = \sum_{J_\beta} \sum_{kl}^{n} \langle \beta(J_\beta)|\hat{E}^\beta_{kl}|\beta(I_\beta)\rangle \left[h_{kl} - \frac{1}{2}\sum_{j}^{n}(kj|jl) \right] c(I_\alpha, J_\beta) + \frac{1}{2}\sum_{J_\beta}\sum_{ijkl}^{n} \langle \beta(J_\beta)|\hat{E}^\beta_{ij}\hat{E}^\beta_{kl}|\beta(I_\beta)\rangle (ij|kl) c(I_\alpha, J_\beta), \quad (113)$$

$$\sigma_2(I_\alpha, I_\beta) = \sum_{J_\alpha} \sum_{kl}^{n} \langle \alpha(J_\alpha)|\hat{E}^\alpha_{kl}|\alpha(I_\alpha)\rangle \left[h_{kl} - \frac{1}{2}\sum_{j}^{n}(kj|jl) \right] c(J_\alpha, I_\beta) + \frac{1}{2}\sum_{J_\alpha}\sum_{ijkl}^{n} \langle \alpha(J_\alpha)|\hat{E}^\alpha_{ij}\hat{E}^\alpha_{kl}|\alpha(I_\alpha)\rangle (ij|kl) c(J_\alpha, I_\beta), \quad (114)$$

$$\sigma_3(I_\alpha, I_\beta) = \sum_{J_\alpha, J_\beta} \sum_{ijkl}^{n} \langle \beta(J_\beta)|\hat{E}^\beta_{ij}|\beta(I_\beta)\rangle \langle \alpha(J_\alpha)|\hat{E}^\alpha_{kl}|\alpha(I_\alpha)\rangle (ij|kl) c(J_\alpha, J_\beta). \quad (115)$$

For efficient implementation, it is convenient to precompute the quantities

$$h'_{kl} = h_{kl} - \frac{1}{2}\sum_{j}^{n}(kj|jl). \quad (116)$$

Note that the first term (σ_1) involves only beta shift operators, the second (σ_2) involves only alpha shift operators, and the third (σ_3) involves both alpha and beta shift operators. These terms are also called the $\beta\beta$, $\alpha\alpha$, and $\alpha\beta$ terms.[48,49] Several determinant-based CI algorithms presented over the last few years[83,183,215,216] have been based on this set of σ equations or the analogous equations for restricted CI (see section 4.8.1).

4.4.2 Simplifications for $M_s = 0$

Certain simplifications arise if the $M_s = 0$ component of an electronic state is used. The first of these is the time-reversal symmetry of the CI vector, which may be expressed as

$$c(I_\alpha, I_\beta) = (-1)^S c(I_\beta, I_\alpha), \quad (117)$$

where S is the spin quantum number. Olsen *et al.* use this fact to show how the σ_2 contribution can be determined entirely from the σ_1 contribution when $M_s = 0$.[46] The remarkably simple result is

$$\sigma_2(I_\alpha, I_\beta) = (-1)^S \sigma_1(I_\beta, I_\alpha). \quad (118)$$

Likewise, it is also possible to show that the $ijkl$-th component of σ_3 satisfies the relation

$$\sigma_3^{ijkl}(I_\alpha, I_\beta) = (-1)^S \sigma_3^{klij}(I_\beta, I_\alpha). \tag{119}$$

This equation may be used to eliminate contributions from $I_\beta > I_\alpha$ or $(kl) > (ij)$, where (ij) and (kl) are compound indices. Olsen argues that the restriction $I_\alpha \geq I_\beta$ is to be preferred where this can be used to eliminate entire blocks of the σ_3 matrix. If all alpha/beta strings with the same irreducible representation are grouped together, then states which are not totally symmetric in their molecular point group will have off-diagonal blocks which can be eliminated using this restriction. On the other hand, when applied to totally symmetric states, this restriction eliminates the upper half of each symmetry block of σ_3. Since this reduces the average vector length in the vectorized algorithm, Olsen recommends using the alternative restriction $(ij) \geq (kl)$ in these cases. This may be accomplished by rewriting σ_3 as

$$\begin{aligned} \sigma_3(I_\alpha, I_\beta) &= \sum_{(ij)\geq(kl)} \sigma_3^{ijkl}(I_\alpha, I_\beta) + \sum_{(ij)<(kl)} (-1)^S \sigma_3^{klij}(I_\beta, I_\alpha) \\ &= \sigma_3'(I_\alpha, I_\beta) + (-1)^S \sigma_3'(I_\beta, I_\alpha), \end{aligned} \tag{120}$$

where

$$\sigma_3'(I_\alpha, I_\beta) = \sum_{(ij)\geq(kl)} \sigma_3^{ijkl}(I_\alpha, I_\beta)(1 + \delta_{(ij),(kl)})^{-1}. \tag{121}$$

Hence the total σ vector can be evaluated as

$$\sigma(I_\alpha, I_\beta) = \sigma_1(I_\alpha, I_\beta) + \sigma_3'(I_\alpha, I_\beta) + (-1)^S \left[\sigma_1(I_\beta, I_\alpha) + \sigma_3'(I_\beta, I_\alpha)\right]. \tag{122}$$

The $M_s = 0$ simplifications therefore reduce computational expense by roughly a factor of two.

One further observation must be made about the loss of spin symmetry in the CI vector in the iterative diagonalization of the Hamiltonian. Even very slight deviations from (117), such as might occur from roundoff errors, become magnified in subsequent iterations and cause the iteration procedure to become numerically unstable because precise adherence to (117) is *assumed* if any of the $M_s = 0$ simplifications just described. If necessary, these difficulties can be avoided by explicitly enforcing the spin symmetry of any new vector in the subspace expansion. In this respect it is important to modify the diagonal elements of the Hamiltonian in the preconditioner for the subspace iteration method, as already discussed in section 3.2.

4.4.3 Algorithms for Computing σ

From (113), one can see that the mathematical operations required to form $\sigma_1(I_\alpha, I_\beta)$ are identical but independent of each other for different I_α. That is,

Figure 7: Olsen's Vectorized Algorithm for σ_1 (Ref. 46).

loop over beta strings I_β
  Zero array F
  Loop over excitations $\hat{E}^\beta_{kl}$ from $|\beta(I_\beta)\rangle$
    $|\beta(K_\beta)\rangle = \mathrm{sgn}(kl)\hat{E}^\beta_{kl}|\beta(I_\beta)\rangle$
    $F(K_\beta) = F(K_\beta) + \mathrm{sgn}(kl)h'_{kl}$
    Loop over excitations $\hat{E}^\beta_{ij}$ from $|\beta(K_\beta)\rangle$
      $|\beta(J_\beta)\rangle = \mathrm{sgn}(ij)\hat{E}^\beta_{ij}|\beta(K_\beta)\rangle$
      $F(J_\beta) = F(J_\beta) + (1/2)\mathrm{sgn}(kl)\mathrm{sgn}(ij)(ij|kl)$
    end loop over $\hat{E}^\beta_{ij}$
  end loop over $\hat{E}^\beta_{kl}$
  loop over beta strings J_β and alpha strings I_α
    $\sigma_1(I_\alpha, I_\beta) = \sigma_1(I_\alpha, I_\beta) + F(J_\beta)c(I_\alpha, J_\beta)$; vect'd
  end loop over I_α, J_β
end loop over I_β

column I_β of σ_1 can be constructed by two multiplications of scalars by columns (J_β) of c. Hence the construction of σ_1 is vectorizable over I_α. The vectorized algorithm for the evaluation of σ_1, adapted from Olsen *et al.*,[46] appears in Figure 7. An analogous algorithm can be used to obtain σ_2. However, one can also obtain σ_2 for $M_s = 0$ cases by (118). These algorithms require the same string replacement lists used by Knowles and Handy[109] (sec. 4.3). Note that the vector F is sparse, and multiplication of F by c should only take place for nonzero values of F.

Unfortunately, the construction of σ_3 (115) is harder to vectorize. A simple, non-vectorized algorithm for σ_3 is presented in Fig. 8. One can see that this does not appear as a simple set of arithmetic operations on vectors. For example, the contributions of the beta strings are not identical for different alpha strings because each alpha string connects to a different set of excited alpha strings with different indices k and l. Olsen *et al.* remedy this by operating a fixed kl at a time;[46] this makes their algorithm vectorizable in the innermost loop. Their algorithm, adapted and expanded from Ref. 46, is presented in Fig. 9. Note that this algorithm also employs scatter/gather (i.e., data rearrangement) operations to ensure that all of the data relevant to the multiplication step $V = Fc'$ are contiguous. This avoids "indirect addressing," which could substantially degrade performance due to long waits for data to be fetched from scattered memory locations.[166] For $M_s = 0$, an improvement to the σ_3 algorithm can be made by utilizing equations discussed in section 4.4.2.

Figure 8: Simple Algorithm for σ_3.

loop over I_α
  loop over $|\alpha(J_\alpha)\rangle = \mathrm{sgn}(kl)\hat{E}^\alpha_{kl}|\alpha(I_\alpha)\rangle$
    loop over I_β
      loop over $|\beta(J_\beta)\rangle = \mathrm{sgn}(ij)\hat{E}^\beta_{ij}|\beta(I_\beta)\rangle$
        $\sigma_3(I_\alpha, I_\beta) = \sigma_3(I_\alpha, I_\beta) + \mathrm{sgn}(ij)\mathrm{sgn}(kl)(ij|kl)c(J_\alpha, J_\beta)$
      end loop over J_β
    end loop over I_β
  end loop over J_α
end loop over I_α

Furthermore, if the integrals possess the full eightfold permutational symmetry, then $\hat{E}^\alpha_{kl}$ can be replaced by $(\hat{E}^\alpha_{kl} + \hat{E}^\alpha_{lk})(1+\delta_{kl})^{-1}$ in order to increase the average vector length in the formation of V. Note once again that F is sparse.

Clearly this algorithm takes less advantage of vector processors than the Knowles-Handy algorithm, since it involves some overhead (setup of the L and R arrays, and the scatter and gather) and uses smaller vector lengths. Nevertheless, one would expect this algorithm to be faster in many cases due to the substantially reduced number of mathematical operations performed. Counting only multiplications, the operation counts for each part of σ are approximately[46]

$$N_1 \approx \frac{1}{4} N_{det} N_\beta^2 (n - N_\beta)^2 \tag{123}$$

$$N_2 \approx \frac{1}{4} N_{det} N_\alpha^2 (n - N_\alpha)^2 \tag{124}$$

$$N_3 \approx N_{det} N_\alpha N_\beta (n - N_\alpha)(n - N_\beta). \tag{125}$$

When $N_\alpha = N_\beta$, the overall operation count is thus approximately

$$N_{op} \approx \frac{3}{2} N_{det} N_\alpha^2 (n - N_\alpha)^2. \tag{126}$$

Recall that this operation count can be cut approximately in half for $M_s = 0$ cases. Knowles and Handy are also able to take advantage of time reversal symmetry for singlet states, by employing the combinations $2^{-1/2}(\alpha(I_\alpha)\beta(I_\beta) + \alpha(I_\beta)\beta(I_\alpha))$. Recalling that the operation count for the Knowles-Handy algorithm is approximately $\frac{1}{4}\tilde{N}_{det} n^4$, we might expect the greatest savings for Olsen's algorithm when n/N is large.

Figure 9: Olsen's Vectorized Algorithm for σ_3 (Ref. 46).

loop over kl
  set up lists $L(I)$, $R(I)$, and $\mathrm{sgn}(I)$, such that
    $|\alpha[L(I)]\rangle = \mathrm{sgn}(I)\hat{E}^{\alpha}_{kl}|\alpha[R(I)]\rangle$
  loop over list entries I and beta strings J_β
    $c'(I, J_\beta) = c(L(I), J_\beta)\mathrm{sgn}(I)$; vect'd gather
  end loop over I and J_β
  loop over I_β
    zero array F
    loop over excitations $\hat{E}^{\beta}_{ij}$ from $|\beta(I_\beta)\rangle$
      $|\beta(J_\beta)\rangle = \mathrm{sgn}(ij)\hat{E}^{\beta}_{ij}|\beta(I_\beta)\rangle$
      $F(J_\beta) = F(J_\beta) + \mathrm{sgn}(ij)(ij|kl)$
    end loop over $\hat{E}^{\beta}_{ij}$
    loop over beta strings J_β and list entries I
      $V(I) = F(J_\beta)c'(I, J_\beta)$; vect'd over I
    end loop over J_β, I
    loop over list entries I
      $\sigma_3(R(I), I_\beta) = \sigma_3(R(I), I_\beta) + V(I)$; vect'd scatter
    end loop over I
  end loop over I_β
end loop over kl

4.5 Zarrabian's Reduced Intermediate Space

Shortly after the publication of the 1988 paper by Olsen *et al.*, Zarrabian, Sarma, and Paldus presented[217] an alternative approach to avoid the unnecessarily large scaling of the Knowles-Handy full CI algorithm. These workers employed an $(N-2)$-electron intermediate space for the two-electron contributions rather than an N–electron intermediate space. Their expressions for σ were originally derived using generators of the group SO(4), but to avoid introducing new notation we will consider the later derivation of Harrison and Zarrabian,[47] which uses only the standard spin-orbital creation and annihilation operators.

We begin by rewriting (24) over spatial orbitals, as

$$\hat{H} = \sum_{ij}^{n} h_{ij} \sum_{\sigma=\alpha,\beta} a_{i\sigma}^{\dagger} a_{j\sigma} + \frac{1}{2} \sum_{ijkl}^{n} (ij|kl) \sum_{\lambda,\sigma=\alpha,\beta} a_{i\sigma}^{\dagger} a_{k\lambda}^{\dagger} a_{l\lambda} a_{j\sigma}. \tag{127}$$

Next, insert the resolution of the identity between the pairs of creation and annihilation operators in the two-electron term. Clearly, the sum must run over $(N-2)$-electron states. The expression for σ becomes

$$\begin{aligned} \sigma_I &= \sum_J H_{IJ} c_J \qquad (128) \\ &= \sum_J \sum_{ij}^{n} h_{ij} \sum_{\sigma=\alpha,\beta} \langle I^{(N)} | a_{i\sigma}^{\dagger} a_{j\sigma} | J^{(N)} \rangle c_J \\ &+ \frac{1}{2} \sum_{JK} \sum_{ijkl}^{n} (ij|kl) \sum_{\lambda,\sigma=\alpha,\beta} \langle I^{(N)} | a_{i\sigma}^{\dagger} a_{k\lambda}^{\dagger} | K^{(N-2)} \rangle \langle K^{(N-2)} | a_{l\lambda} a_{j\sigma} | J^{(N)} \rangle c_J, \end{aligned}$$

where the superscripts (N) and $(N-2)$ denote the number of electrons for each state. The one-electron terms are exactly the same as before. The two-electron contributions to σ, denoted $\sigma^{(2)}$, may be written in terms of separate contributions from each possible spin case:

$$\begin{aligned} \sigma_{\alpha\alpha}^{(2)} &= \sum_{JK} \sum_{i>k,j>l}^{n} [(ij|kl) - (il|jk)] \langle I^{(N)} | a_{i\alpha}^{\dagger} a_{k\alpha}^{\dagger} | K^{(N-2)} \rangle \langle K^{(N-2)} | a_{l\alpha} a_{j\alpha} | J^{(N)} \rangle c_J \\ \sigma_{\beta\beta}^{(2)} &= \sum_{JK} \sum_{i>k,j>l}^{n} [(ij|kl) - (il|jk)] \langle I^{(N)} | a_{i\beta}^{\dagger} a_{k\beta}^{\dagger} | K^{(N-2)} \rangle \langle K^{(N-2)} | a_{l\beta} a_{j\beta} | J^{(N)} \rangle c_J \\ \sigma_{\alpha\beta}^{(2)} &= \sum_{JK} \sum_{ijkl}^{n} (ij|kl) \langle I^{(N)} | a_{i\beta}^{\dagger} a_{k\alpha}^{\dagger} | K^{(N-2)} \rangle \langle K^{(N-2)} | a_{l\alpha} a_{j\beta} | J^{(N)} \rangle c_J, \end{aligned} \tag{129}$$

where the restrictions over the orbital indices in the $\sigma_{\alpha\alpha}^{(2)}$ and $\sigma_{\beta\beta}^{(2)}$ terms is made possible by the permutational symmetry of the integrals and the anticommutation relations of the creation and annihilation operators. Likewise, the two

Figure 10: Harrison and Zarrabian's Vectorized Algorithm for $\sigma^{(2)}_{\alpha\alpha}$ (Ref. 47).

loop over alpha strings I_α
  loop over orbital pairs $i > k$ (creation op.)
    define $(N_\alpha - 2)$-electron string K_α
    $|\alpha(K_\alpha)\rangle = \mathrm{sgn}(ik) a_{k\alpha} a_{i\alpha} |\alpha(I_\alpha)\rangle$
    loop over orbital pairs $j > l$ (annihilation op.)
      define new N_α-electron string J_α
      $|\alpha(J_\alpha)\rangle = \mathrm{sgn}(jl) a^\dagger_{j\alpha} a^\dagger_{l\alpha} |\alpha(K_\alpha)\rangle$
      $V = \mathrm{sgn}(ik)\mathrm{sgn}(jl)[(ij|kl) - (il|kj)]$
      loop over beta strings I_β
        $\sigma^{(2)}_{\alpha\alpha}(I_\beta, I_\alpha) = \sigma^{(2)}_{\alpha\alpha}(I_\beta, I_\alpha) + V * c(I_\beta, J_\alpha)$

mixed contributions $\alpha\beta$ and $\beta\alpha$ have been combined in $\sigma^{(2)}_{\alpha\beta}$, eliminating the coefficient of $\frac{1}{2}$.

The algorithm for constructing $\sigma^{(2)}_{\alpha\alpha}$, adapted from Harrison and Zarrabian,[47] is given in Fig. 10. The algorithm for $\sigma^{(2)}_{\beta\beta}$ is of course analogous. It is easy to show that the number of floating-point multiplications involved in the construction of $\sigma^{(2)}_{\beta\beta}$ and $\sigma^{(2)}_{\alpha\alpha}$ with this algorithm are

$$N_1 = \frac{1}{4} N_{det} N_\beta (N_\beta - 1)(n - N_\beta + 2)(n - N_\beta + 1) \tag{130}$$

$$N_2 = \frac{1}{4} N_{det} N_\alpha (N_\alpha - 1)(n - N_\alpha + 2)(n - N_\alpha + 1), \tag{131}$$

which are basically the same as the approximate operation counts (123)-(124) for Olsen's algorithm.[46] Harrison and Zarrabian point out that this algorithm can be parallelized over the outermost loop. Note that they address the CI vector with the beta string as the row index instead of the alpha string; the earlier paper by Zarrabian *et al.* used the alternative convention. This choice can have some relevance for $\sigma^{(2)}_{\alpha\alpha}$ and $\sigma^{(2)}_{\beta\beta}$ when only one of the terms is explicitly constructed (i.e., when $M_s = 0$). In that case, it is best to access the data in C sequentially (i.e., with "unit stride").[166]

For $\sigma^{(2)}_{\alpha\beta}$, one can use a similar loop structure to that in Fig. 9 or Fig. 10.[217] This yields an operation count[47] of

$$N_3 = N_{det} N_\alpha N_\beta (n - N_\alpha + 1)(n - N_\beta + 1). \tag{132}$$

However, Harrison and Zarrabian suggest that for parallel-vector machines, it is better to revert to a matrix multiplication such as that used by Knowles and Handy.[109] This algorithm is produced in Fig. 11. These loops are run

Figure 11: Harrison and Zarrabian's Vectorized Algorithm for $\sigma^{(2)}_{\alpha\beta}$ (Ref. 47).

precompute info for adding orbs to $(N_\alpha - 1)$-elec. α string
precompute info for adding orbs to $(N_\beta - 1)$-elec. β string
zero D
loop over orbitals l to be added to $(N_\alpha - 1)$-elec. string K_α
  loop over orbitals j to be added to $(N_\beta - 1)$-elec. string K_β
    loop over $(N_\alpha - 1)$-electron strings K_α
      define N_α–electron string J_α
      $|\alpha(J_\alpha)\rangle = \mathrm{sgn}(l) a^\dagger_{l\alpha} |\alpha(K_\alpha)\rangle$
      loop over $(N_\beta - 1)$-electron strings K_β
        define N_β–electron string J_β
        $|\beta(J_\beta)\rangle = \mathrm{sgn}(j) a^\dagger_{j\beta} |\beta(K_\beta)\rangle$
        $D(K_\beta, K_\alpha, jl) = D(K_\beta, K_\alpha, jl) + \mathrm{sgn}(k)\mathrm{sgn}(j) c(J_\beta, J_\alpha)$
      end loop over K_β
    end loop over K_α
  end loop over j
end loop over l

call optimized matrix multiply for $E_{K,ik} = D_{K,jl}\langle jl|ik\rangle$

loop over orbitals k to be added to $(N_\alpha - 1)$-elec. string K_α
  loop over orbitals i to be added to $(N_\beta - 1)$-elec. string K_β
    loop over $(N_\alpha - 1)$-electron strings K_α
      define N_α–electron string I_α
      $|\alpha(I_\alpha)\rangle = \mathrm{sgn}(k) a^\dagger_{k\alpha} |\alpha(K_\alpha)\rangle$
      loop over $(N_\beta - 1)$-electron strings K_β
      $|\alpha(I_\alpha)\rangle = \mathrm{sgn}(k) a^\dagger_{k\alpha} |\alpha(K_\alpha)\rangle$
        define N_β–electron string I_β
        $|\beta(I_\beta)\rangle = \mathrm{sgn}(i) a^\dagger_{i\alpha} |\beta(K_\beta)\rangle$
        $\sigma^{(2)}_{\alpha\beta}(I_\beta, I_\alpha) = \sigma^{(2)}_{\alpha\beta}(I_\beta, I_\alpha) + \mathrm{sgn}(i)\mathrm{sgn}(k) E(K_\beta, K_\alpha, ik)$
      end loop over K_β
    end loop over K_α
  end loop over i
end loop over k

for blocks of several intermediate states K at a time, and additional loops account for spatial symmetry. Note the use of two-electron integrals in Dirac notation rather than Mulliken notation; i.e., $\langle jl|ik\rangle = (ij|kl)$ for real integrals. Furthermore, the integrals are stored without any permutational symmetry. This algorithm has an operation count of

$$N_3 = N_{det}N_\alpha N_\beta n^4/(n - N_\alpha + 1)(n - N_\beta + 1), \tag{133}$$

which can be obtained[47] by using $N_\alpha/(n-N_\alpha+1)$ as the ratio of the number of $(N_\alpha - 1)$-electron strings to the number of N-electron strings (and by ignoring spatial symmetry). Note that this operation count is not too much greater than that of the non-matrix version (132) when $m >> N_\alpha, N_\beta$. In such cases, and given $N_\alpha = N_\beta$, the overall number of multiplications $N_1 + N_2 + N_3$ is about $\frac{3}{2}N_{det}N_\alpha^2 n^2$, compared to $\frac{1}{4}\tilde{N}_{det}n^4$ for the Knowles-Handy algorithm.

Although the work done by this algorithm is basically equivalent to that done in Olsen's algorithm, Zarrabian *et al.* suggest[47,217] that their approach would be better suited for the evaluation of three- and four-electron reduced density matrices, which are important in the context of internally contracted MR-CISD.[101] They also note that it should be possible to adapt their algorithm to restricted CI spaces,[47,217] and some work along these general lines has been presented by Duch.[203]

4.6 The Table-Based Algorithm of Bendazzoli and Evangelisti

Using Handy's alpha and beta string formalism,[44] along with some of the notation of Olsen,[46] Bendazzoli and Evangelisti have presented a full CI algorithm[48,49] which uses tables to represent the excitation operators $\hat{E}_{ij}^{\beta}$ rather than the string replacement lists of Knowles and Handy. The operation count of their method is essentially the same as that of Olsen *et al.*[46] and of Zarrabian *et al.*,[47,217] but the data are organized differently and the authors note that their loop structure is more easily parallelized than that of Olsen *et al.*[46] The algorithm of Bendazzoli and Evangelisti[48] for σ_1 (which they call the $\beta\beta$ term), is presented in Figure 12. When $M_s = 0$, σ_2 can be obtained from (118) just as in Olsen's approach.[46]

The tables $OOVV$ represent the shift operator products $\hat{E}_{il}^{\beta}\hat{E}_{jk}^{\beta}$; for a given set of orbitals (i, j, k, l), $OOVV(i, j, l, k)$ gives a list of all beta strings with orbitals i, j occupied and l, k unoccupied. This is the list of all strings which can be acted on to the left by the shift operator product. Similarly, $OOVV(l, k, i, j)$ gives a list of all strings which can be acted on to the right by this same product. The clever aspect of this approach is that the Ith entry of $OOVV(i, j, l, k)$ (denoted I_1) is the *same* as the string produced by applying $\hat{E}_{il}^{\beta}\hat{E}_{jk}^{\beta}$ to the Ith

Figure 12: Bendazzoli and Evangelisti's Algorithm for σ_1 (Ref. 48).

loop over $i > j, k > l$
  $V = (ik|jl) - (il|jk)$
  loop over $I = 1$, length of list $OOVV(i,j,k,l)$
    $I_1 = I$th entry of list $OOVV(i,j,l,k)$
    $S_1 =$ sign associated with I_1
    $VS = V * S_1$
    $I_2 = I$th entry of list $OOVV(l,k,i,j)$
    loop over $J = 1$, number of alpha strings
      $\sigma(J,I_2) = \sigma(J,I_2) + c(J,I_1) * VS$
    end loop over J
  end loop over I
end loop over i,j,k,l

element of $OOVV(l,k,i,j)$ (denoted I_2). Bendazzoli and Evangelisti have so far limited their attention to full CI; for restricted CI, the size of the lists OV and $OOVV$ will rapidly become large relative to the size of the CI vector (sec. 4.9.4), so that these lists are probably appropriate only for full CI.

The σ_3 algorithm is presented in Figure 13. Note the same scatter/gather structure as in Figures 9 and 22. Like our own version (cf. section 4.9.5), this algorithm eliminates the F array and uses a DAXPY operation[166] in the innermost loop. Compared to the algorithm in Figure 22, our initial attempts to implement this algorithm for σ_3 yielded a program running roughly 50% slower on the IBM RS/6000 POWER2 model 3CT workstation.

More recently, Evangelisti, Bendazzoli, and co-workers have developed a parallel implementation of their algorithm for the Cray T3D, a distributed memory machine.[50,84] The newest out-of-core version of their program allows the CI and σ vectors to be processed one symmetry block at a time. To avoid storage of the diagonal of the Hamiltonian, they approximate it using orbital eigenvalues. Following Olsen[83] (sec. 3.2.2), they minimize storage space by using only one CI vector and one σ vector in their iterative diagonalization method, although the details of their iterative procedure differ somewhat from those of Olsen and co-workers. In 1996, this parallelized version was used on a 64-processor Cray T3D to obtain[84] the full CI wavefunction for Be_2, with all electrons correlated and using a 9s2p1d basis (derived from a 4s2p1d ANO basis by uncontracting the primitive Gaussians corresponding to the five largest coefficients in the first ANO orbital). This represents the first converged CI calculation requiring more than one billion Slater determinants

Figure 13: Bendazzoli and Evangelisti's Algorithm for σ_3 (Ref. 48).

loop over k, l

  loop over $I = 1$, length of list $OV(l, k)$
    $I_2 = I$th entry of $OV(l, k)$
    $S_2 =$ sign associated with I_2
    loop over $J = 1$, number of beta strings
      $c'(I, J) = c(I_2, J) * S_2$
    end loop over J
  end loop over I

  loop over i, j
    $V = (ij|kl)$
    loop over $J = 1$, length of list $OV(i, j)$
      $J_1 = J$th entry of $OV(i, j)$
      $J_2 = J$th entry of $OV(j, i)$
      $S_2 =$ sign associated with J_2
      $VS = V * S_2$
      loop over $I = 1$, length of $OV(l, k)$
        $\sigma'(I, J_2) = \sigma'(I, J_2) + c'(I, J_1) * VS$
      end loop over I
    end loop over J
  end loop over i, j

  loop $I = 1$, length of list $OV(l, k)$
    $I_1 = I$th entry of $OV(l, k)$
    loop $J = 1$, number of beta strings
      $\sigma(I_1, J) = \sigma'(I, J)$
    end loop over J
  end loop over I

end loop over k, l

(an unconverged calculation on the Mg atom involving more than a billion determinants was reported in 1990 by Olsen, Jørgensen, and Simons[83]).

4.7 Approximate Full CI Methods

In 1989, Knowles introduced[81] a modified full CI procedure which exploits the sparsity of the Hamiltonian matrix and affords approximate full CI results at a dramatically reduced computational cost. Employing the Davidson method[108] (cf. section 3.2.1), the correction to the current CI vector is given by

$$\Delta c_I = \frac{r_I}{(E - H_{II})}, \tag{134}$$

where r_I is the residual $\sigma_I - Ec_I$. Knowles estimates the importance of these corrections using the following simple expression inspired by second-order perturbation theory:

$$\Delta E_I = r_I \Delta c_I. \tag{135}$$

If $|\Delta E_I|$ is less than some threshold, Δc_I is neglected. Thus far fewer determinants are actually included in the correction vector, which is stored on disk in a packed format. One problem with this approach is that neglected corrections Δc_I can reappear during the standard Schmidt orthogonalization against previous subspace vectors (cf. section 3.2). Knowles thus avoids the Schmidt orthogonalization step and employs a non-orthogonal space of expansion vectors. This allows for tight control over the size of the expansion space vectors.

A potential difficulty of this approach is that the σ vectors (which must also be stored) are not necessarily sparse. Knowles notes[81] that even when $\mathbf{c}$ is only 1% populated, typically 50% of σ will be nonzero. Nevertheless, in order to obtain variationally correct energies, the full σ vector must be formed in core memory and its dot product taken with all expansion vectors $\mathbf{c}$. However, once this is done, the only further use of σ is in the construction of new subspace vectors; hence, Knowles only writes to disk those elements of σ greater than some threshold. According to (134)-(135), these neglected elements of σ would only contribute to elements of $\Delta\mathbf{c}$ which make very small energy contributions.

Given the Knowles-Handy full CI algorithm of section 4.3, it is clear that the matrix formulation no longer applies with a sparse CI vector $\mathbf{c}$. Instead, the the formation of σ is driven from the list of nonzero elements in $\mathbf{c}$, employing scatter and gather operations to obtain some vectorization in the innermost loops; this approach is therefore similar to the original string-driven approach of Handy[44] or the subsequent algorithm of Olsen *et al.*[46] To avoid core storage problems, the exact σ can be formed one symmetry block at a time (where a symmetry block of σ contains all elements σ_I having the same alpha string

symmetry). Memory requirements can be further reduced, with some loss in efficiency, by processing σ in smaller batches of arbitrary size.[81]

Knowles and Handy demonstrated the power of this approach by estimating the full CI energy of NH_3 in an atomic natural orbital (ANO) basis set of DZP quality.[218] The full CI expansion contains more than 209 million determinants, yet Knowles and Handy were able to obtain an apparently reliable variational energy of -56.4235 hartree using a CI vector with only 665,247 nonzero elements (0.3% of the full CI vector). Employing perturbation theory to estimate the remaining energy error (presumably via equation 135), Knowles and Handy arrived at a final estimate of -56.4236 $\pm$ 0.0001 hartree.[218]

Using perturbation theory to estimate the importance of determinants in configuration interaction is a very old idea (see Ref. 57 for a detailed review). Indeed, it is perturbation theory which provides the justification for truncating the CI space at only singles and doubles from one or several references (i.e., the CISD and MR-CISD methods). The CIPSI method (1973) of Huron, Malrieu, and Rancurel[92] diagonalizes the Hamiltonian in some subspace of selected determinants and uses the resulting eigenvector as the zeroth-order wavefunction in a subsequent perturbation theory treatment. Determinants having a contribution to the first-order wavefunction greater than some threshold η are added to the selected CI space, and this process is repeated until the selection threshold is considered acceptably small or until the selected CI space becomes too large to handle. The effect of unselected determinants is evaluated by second-order perturbation theory. The procedure of Knowles[81,218] is similar to this, but differs in two important respects: first, Knowles selects determinants based on a perturbative estimate of their contribution to the energy rather than to the first-order wavefunction, and second, Knowles applies the selection *during* the Davidson procedure, whereas CIPSI solves the CI problem exactly for each selected CI space. A more recent version of the CIPSI method[93] is somewhat more flexible and introduces a third class of determinants of intermediate importance; interacting determinants with an estimated CI coefficient less than η but greater than a second threshold τ can be treated by higher-order perturbation theory or variationally, while those with contributions less than τ are treated by second-order perturbation theory as before. The CIPSI scheme should yield wavefunctions approaching the full CI limit, and indeed it has been benchmarked against full CI.[93,122,123,215,219] The most recent studies have added a self-consistent dressing of the Hamiltonian matrix to ensure size consistency.[122,123]

Another long-established approach to approximating full CI is to employ successively larger MR-CISD spaces. Since the size of the CI space grows very rapidly as the number of references is increased, Buenker and Peyerimhoff (1974-5)[10,11] suggested retaining only the most important singly and doubly

substituted configurations and treating discarded configurations by Brillouin-Wigner perturbation theory and extrapolation procedures; they call their procedure MRD-CI. Their strategy implies that the most compact wavefuntions are obtained by truncating the singles and doubles space rather than the reference space, and indeed CIPSI studies support this idea.[122,219] Unlike the CIPSI method, Buenker and Peyerimhoff do not use perturbation theory in the configuration selection; rather, orbital configurations are accepted or rejected on the basis of the energy lowering they cause when added to the reference space. A separate small CI procedure is required for each possible spatial orbital configuration. Although this may require somewhat more effort than the perturbational estimates of CIPSI, the energy lowerings can be reused in the extrapolations to zero threshold.[10,11] Alternative approaches to making MR-CISD more computationally tractable are the internal and external contraction schemes discussed in section 2.4.2.

Knowles' 1989 program[81,218] was able to approach the full CI limit more closely than selected CI methods such as MRD-CI and CIPSI because it was efficient enough to treat a much larger number of determinants variationally. Subsequently in 1992, Povill, Rubio, and Illas noted[215] that the principal difficulty with the standard CIPSI program was its need to store the Hamiltonian matrix **H**, allowing it to handle no more than 50,000 determinants variationally. Hence, they presented[215] the direct selected configuration interaction using strings (DISCIUS) algorithm employing the alpha and beta string formalism of Handy;[44] the notation, $(\alpha\alpha, \beta\beta, \alpha\beta)$ spin decomposition, and $M_s = 0$ simplifications of Olsen *et al.*;[46] and the reduced intermediate space of Zarrabian *et al.*[47,217] Special ordering and addressing schemes, which make use of large index arrays, allow for some degree of vectorization despite the lack of a well-defined structure in the CI space.[215] Nevertheless, a more recent (1995) version of this algorithm by Povill and Rubio[220] largely abandons the vectorization of σ_3, noting that the average vector length for selected CI spaces is generally too small for effective vectorization. These authors also found that too much time is spent checking to see if doubly excited strings in the construction of σ_1 or σ_2 belong to the selected space; hence, they consider every pair of allowed strings and determine all single and double excitations connecting them. The DISCIUS algorithm is capable of treating selected CI spaces with more than one million determinants.[220]

A related algorithm, which has also been coupled to the CIPSI approach, was presented by Caballol and Malrieu[221] in 1992. Their approach is also direct and determinant-based, but the strings are written as particle-hole excitations from a single reference state; the program is named SCIEL, for selected CI with excitation labeling. For a determinant with excitation level m, the particle-hole labeling lists m holes and m particles. This is inefficient for full CI,[221] since

it would require the listing of $2N_\alpha$ orbitals for a maximally-excited ($m = N_\alpha$) alpha string, rather than only N_α orbitals in the standard approach. However, for CI spaces dominated by determinants with a relatively low excitation level, this formalism could offer some benefits. Povill *et al.* have commented that the DISCIUS and SCIEL programs seem to have similar efficiencies.[122]

Similar improvements have been made to the MRD-CI program of Buenker and Peyerimhoff,[10,11] which was previously limited to about 50,000 configurations.[222] In 1995, Krebs and Buenker presented[222] a new table-direct CI algorithm for use in the MRD-CI selection scheme which is capable of handling variational spaces including at least several hundred thousand determinants.

Knowles' 1989 sparse CI method has been the subject of additional study in the last few years. In 1994, Mitrushenkov presented a very similar method[183] which differs primarily in that it selects components of the CI vector based on their magnitude (134) and not on their expected energy lowering (135). This choice was motivated by the belief that it would yield more physical CI vectors less likely to give errors for properties other than the total energy.[183] Mitrushenkov described how to adapt Olsen's full CI algorithm to implement his approach, which he has called dynamic CI. Of particular interest is his technique for avoiding core storage of the entire σ vector: he calculates $\sigma(I_\alpha, I_\beta)$ for a fixed I_β (i.e., the algorithm is driven by σ rather than by nonzero elements of **c**). The exact σ values are used to update the Hamiltonian in the small Davidson subspace, and then components larger than a given cutoff are written to disk. Like Knowles, Mitrushenkov uses a nonorthogonal Davidson subspace; however, he uses only two vectors and employs the improved preconditioner of Olsen *et al.*[83] (cf. section 3.2.2). Mitrushenkov reported results for NH_3, H_2O, and Mg test cases,[183] but unfortunately no results were presented for systems where the exact full CI result was known (DZP NH_3 full CI results have subsequently been reported,[50,80] see below).

In 1991, Harrison emphasized the use of second-order perturbation theory to approach the full CI results more rapidly.[223] In Harrison's method, denoted CI+PT, one chooses an initial reference space (perhaps a single determinant), and an initial selection threshold η. A CI is performed in the reference space, yielding eigenvectors for all roots of interest. Unlike most of the other algorithms discussed in this section, Harrison's program employs CSFs rather than determinants; two-electron coupling coefficients are evaluated as products of one-electron coupling coefficients, which in turn are evaluated by the method of Knowles and Werner[102] (see section 4.1). For every configuration I interacting with (but not included in) the reference space, the second-order perturbation theory energy contribution is determined for each of the desired roots k, using