An Implementation of Configuration Interaction in a General Purpose Semiempirical Context

Daniel A. Liotard*,† and Andrew Holder‡

Laboratoire de Physico-Chimie Theorique, University Bordeaux I, 351 Cours de la Liberation 33405, Talence France, and Department of Chemistry, University of Missouri−Kansas City, 5009 Rockhill Road, Kansas City, Missouri 64110

Received October 2, 1998

We illustrate here an efficient approach to the implementation of configuration interaction (CI) methodologies in the case of general purpose semiempirical Hamiltonians such as MINDO/3, MNDO, PM3, AM1, and SAM1. The strategy described here has been implemented in the commercial program AMPAC and successfully blends well-known and tested algorithms. The primary purpose is to derive a reliable method for determining a CI solution with minimal user input. The success and intrinsic accuracy of the approach is discussed.

I. INTRODUCTION

General-purpose semiempirical methods such as MINDO3,[1] MNDO,[2] AM1,[2] PM3,[4] MNDO/d,[5] and SAM1[6] are primarily concerned with the prediction of ground state molecular properties at equilibrium geometries. This is where the methods were parametrized and where the vast preponderance of data exists. The theoretical approach used for this purpose is almost invariably the restricted or unrestricted Hartree−Fock (RHF/UHF) self-consistent field (SCF) theory, a level that has proved sufficient to accurately reproduce the properties of "ordinary" organic molecules.[7]

However, there are several cases where this first-level approximation is not applicable. Perhaps the most obvious use is the description of excited states for the prediction of UV/visible spectra. The need for efficient semiempirical methods capable of calculating these quantities has long been recognized, especially in cases of larger molecules. Dedicated methods parametrized specifically for this application have been proposed and implemented.[8,9] These approaches are based on a post-SCF configuration interaction (CI) variational formalism restricted to single excitations (SCF-CIS). A variation-perturbation treatment of primarily single and double excitations was used in a PERTCI approach for many years[10] and was effective at reproducing UV−visible spectra of molecules in a semiempirical context.

One of the other situations where SCF alone does not offer sufficient flexibility include the treatment of certain transition states corresponding to (weakly) avoided crossings. Other examples are systems that are *not* closed shell singlet in the ground state, mostly involving transition metal complexes or atoms. In both cases, the UHF-SCF[11] and/or the "half-electron"[12] formalisms may improve the results but at a cost. UHF/SCF introduces spin contamination to the final solution, bringing into question the overall quality of the results, since the spin state is not formally correct. The half-electron (or more properly "partial-electron" method for spin states higher than a doublet) method not only deviates from the formal requirements of self-consistency but also delivers a *mean* energy for the spin states implicitly involved. Again, for these cases a variational CI treatment must be invoked to deliver the exact energy for the spin multiplicity of the particular state under study. Thus, although drastically limited by combinatorial expansion, this CI does not reduce to a CI with single excitations only (CIS) when numerous open shells are implied as do the previously mentioned INDO methods. It is thus apparent that modern multipurpose semiempirical packages must offer a CI capability that is not restricted to either single (CIS) or single/double excitations (CISD) only.

Beginning from the basic RHF-SCF formalism, a reasonable and useful set of molecular orbitals (MO) can be generated that will eventually contain unpaired electrons. Following this, a natural approach to CI is to construct a complete active space (CAS) variational treatment and then delete some of the states from consideration so that a limited number of important and contributory states (including the ones that end up in the unpaired description) remain for the CI matrix diagonalization. Thus, such a CAS-CI method has the potential to successfully describe ground and first excited states (for spectroscopy and photochemistry), transition states, and systems with numerous open shells.

A CAS-CI treatment was introduced quite some time ago in well-known semiempirical packages.[13] But these implementations suffer from the inherent combinatorial explosion accompanying the increase in the size of the active space needed for effective description of the species of interest. The problem is crucial in a semiempirical context when preliminary calculations such as integrals, SCF, and four-index transformation of the two-electron integrals are very fast with respect to the CI matrix diagonalization step. This occurs when the CI active space reaches seven MOs, while for problems of more general interest 15−25 MOs are required.

† University Bordeaux I. Phone: (33) 556 84 63 20; e-mail: liotard@frbdx11.cribx1.u-bordeaux.fr.

‡ University of Missouri−Kansas City. Phone: (816) 235-2293; e-mail: aholder@cctr.umkc.edu.

10.1021/ci980149i CCC: $18.00
Published on Web 05/06/1999

Numerous methods have been proposed to handle the reduction of states considered in CI, but these have mostly been applied in an *ab initio* MO context, not semiempirical. This paper describes the details of method and calibration of techniques tuned for best performance with semiempirical methods and implemented in the general-purpose semiempirical package AMPAC.[14] The strategy of diagonalization is described in Section II, the procedure to generate, sort, and select the determinants is given in Section III, and the confidence in truncation of the series of determinants is studied in Section IV.

II. DIAGONALIZATION PROCEDURE

Preliminary tests and computational contingencies indicate that the CI matrix to be diagonalized can often reach a size of hundreds of states but should not usually exceed a size of several thousand states. Matrices of these sizes are accessible to direct diagonalization methods as QR[15] but at a cost rapidly prohibitive as the size increases. Common alternatives include indirect methods such as the Lanczos[16] or Davidson[17] classes of algorithms. As the objective after determinant sorting and selection is to retain just a few states of the lowest eigenvalues and eigenvectors, we focused on the Davidson procedure, which is generally recognized to be better suited for the present case than Lanczos's method, as it takes better advantage of the initial guess.

Let **A** be the real, symmetric matrix of size *M* to diagonalize and let **B** be a basis set of orthonormalized column vectors spanning a "model" space of moderate size *m*. The target eigenset **C** is of size *l*, with *l* less than or equal to *m*. A Davidson strategy reduces to the following.

Step 1A. Projection onto the model space and direct diagonalization

$$(\mathbf{B}'\mathbf{A}\mathbf{B})\mathbf{V} = \mathbf{V}\mathbf{E}$$

where **V** is a matrix of size *m* containing the orthogonal eigenvectors and **E** is a diagonal matrix of size *m* of the eigenvalues in algebraic ascending order. Current approximation for the target is $\mathbf{C} = \mathbf{B}\mathbf{V}$ and **E**.

Step 1B. (Optional) Replace **B** with **C** and *m* is decreased to *l*.

Step 2. Residual column vectors $\mathbf{R}_i$ (*i* from 1 to *l*) are constructed:

$$\mathbf{R}_i = \mathbf{A}\mathbf{B}\mathbf{V}_i - \mathbf{E}_i\mathbf{B}\mathbf{V}_i$$

Step 3. Construct complementary model vectors **G** from a generator operator **O**:

$$\mathbf{G}_i = \mathbf{O}_i\,\mathbf{R}_i \quad (i \text{ from 1 to } l)$$

Step 4. Orthonormalize **G** to **B**, yielding *k* independent new vectors **N** ($1 \le k \le l$). Append **N** to **B** and return to step 1 with *m* increased by *k*.

Ideally, step 1B is bypassed and the size *m* of the model space increases at each cycle, ensuring convergence at the start of any guess **B**. Practically, step 1B reduces *m* conveniently while retaining most of the information gathered so far. However, the convergence is no longer assured unless special conditions are met by the matrix **A** and the generator operator **O**. This is described below.

The basic implementation of the method searches for only one eigenvector at a time. When coupled with additional orthonormality conditions, the eigenvectors are encountered in increasing order, and the deflation procedure is expensive. Since the size *M* of the matrix **A** is not very large in the present case, there is no computational limit to a simultaneous search for the entire eigenset of size *l*. Substantial savings result. The procedure is considered to be convergent when the change in **E** and the residues **R** fall below prescribed tolerances. These values are typically 0.000 000 1 eV on **E** and 0.0001 eV on the **L1** norm of **R**.

The choice for the generator operator **O** is the key for success of the method and has been recently scrutinized.[18] The matrix **O** may be the unit matrix **I**, yielding (in the absence of step 1B) a strict relaxation (steepest descent behavior) behavior:

$$\mathbf{O} = \mathbf{I} \tag{1}$$

Other forms of the operator **O** attempt to retain a majority of the quadratic behavior as included in the full inverse power method:[19]

$$\mathbf{O}_i = (\mathbf{A} - \mathbf{E}_i\mathbf{I})^{-1} \tag{2a}$$

The matrix inversion appearing in eq 2a rapidly becomes intractable, and the right member has to be replaced by the first term of a more suitable expansion. Partitioning the matrix **A** into two segments **T** and **U**, where $\mathbf{A} = \mathbf{T} + \mathbf{U}$, yields

$$(\mathbf{A}-\mathbf{E}_i\mathbf{I})^{-1} = (\mathbf{T}-\mathbf{E}_i\mathbf{I})^{-1} - (\mathbf{T}-\mathbf{E}_i\mathbf{I})^{-1}\mathbf{U}(\mathbf{T}-\mathbf{E}_i\mathbf{I})^{-1} + \ldots \tag{2b}$$

The matrix **T** must be easy to invert for the sake of computational feasibility. Davidson's original choice takes $\mathbf{T} = \mathbf{D}$ (where **D** is the diagonal of **A**) and retains only the first term of the expansion, yielding a standard perturbative correction to the eigenvectors:

$$\mathbf{O}_i = (\mathbf{D}-\mathbf{E}_i\mathbf{I})^{-1} \tag{3}$$

With this choice for the matrix **T**, the necessary and sufficient condition of convergence of the expansion in Equation 2b is that all eigenvalues of the non-Hermitian matrix $\mathbf{U}(\mathbf{D}-\mathbf{E}_i\mathbf{I})^{-1}$ are less than one in modulus. This property is not usually met for CI matrices, and step 1B is therefore to be avoided as much as possible. This highlights the difficulty in the choice of the first guess for **B** as was pointed out in the literature.[18] In tests using this method, we actually observed chaotic behaviors, making this method of solution unacceptable for automated use, since optimizations of geometry, walks on potential energy surfaces (PES), and other tasks require a very accurate calculation of the energy at each point to be useful.

In our applications, the CI problem is not expressed in operator form, but the matrix **A** is known explicitly. This aspect of our formulation makes attractive an 1969 proposal by Weltin[20] to partition the matrix **A** in its lower left corner (diagonal included) **T** and the remaining portion **U** yielding to

$$\mathbf{O}_i = (\mathbf{T}-\mathbf{E}_i\mathbf{I})^{-1} \tag{4}$$

Table 1. Number of Cycles and Timings (s) of Generator Operators as Functions of Rules B and C, with Rule A always "on"

rules/operators		eq 1		eq 3		eq 4	
rule B	rule C	cycles	time	cycles	time	cycles	time
off	off	53	(51.8)	29	(35.8)	34	(37.9)
off	on	62	(35.2)	36	(31.1)	33	(31.3)
on	off	28	(46.7)	25	(43.8)	17	(43.1)
on	on	32	(46.5)	30	(43.0)	18	(39.9)

where

$$\mathbf{T}_{kl} = \mathbf{A}_{kl}, \quad \mathbf{U}_{kl} = 0 \quad \text{when } k \geq l$$

$$\mathbf{T}_{kl} = 0, \quad \mathbf{U}_{kl} = \mathbf{A}_{kl} \quad \text{when } k < l$$

This choice for the matrix **T** ensures that the convergence of the expansion (Equation 2b) under the mild requirement that the matrix $(\mathbf{D}-\mathbf{E}_i\mathbf{I})$ be definite (Nekrasov's theorem[21]). The generator operator in Equation 4 is just the first step in a Gauss-Seidel procedure to solve a linear system of equations. The robustness of this strategy of mixing Davidson's basic scheme and Weltin's single iteration method appears to be quite successful. This is evidenced by the disappearance of chaotic behaviors using the following rules of thumb.

Rule A. For the very first cycles, adopt a value for *l* larger than the targeted value. This value should be as large as possible at the beginning and then halved at each new cycle until the target value is reached.

Rule B. In Step 1B, adopt a value for *m* a bit larger than *l*, (typically $1.33 \times l$).

Rule C. Perform Step 1B when *m* exceeds $3 \times l$.

Rules A and B improve robustness and stability in the diagonalization. Rule B also allows detection of a degeneracy of states at the boundary *l*, either accidental or imposed by the spatial symmetry point group. Rule C reduces the computational effort and thus the expense.

Because any relaxation progress (at Step 3) depends on information that progressively disappears as it is approached (at Step 2), care must be taken in the orthonormalization at Step 4. The Schmidt procedure is not regular and well-behaved, often producing after a few cycles a basis set **B** that is not orthogonal when the size *l* of the target is large. Liotard and Dewar have demonstrated a much more robust procedure replacing the Schmidt orthogonalization with the hierarchical orthogonalization[22] that has proved to be efficient in solving coupled perturbative Hartree−Fock equations by Patchkovskii and Thiel.[23] The technique's performance is competitive with previously described approaches.[24] The additional cost of diagonalization of the matrix $(\mathbf{B}+\mathbf{G})'$-$(\mathbf{B}+\mathbf{G})$ is balanced by a slower extension of the size *m* of the model space (as typically $k/l < 0.4$), requiring less frequent invocation of the potentially troublesome Step 1B.

The first guess for **B** comes from a direct diagonalization of the upper left corner (ULC) of the matrix **A** (the dominant configurations), of size nearly $3 \times l$. This step is not expensive as some sort of the dominant configurations is part of the selection of determinants described below in Section III.

Table 1 shows the performance of the three generator operators in the case of a typical CI matrix of order $M = 1000$, with $l = 30$ as a target. Several eigenvectors have significant but not pathologically large components out from the ULC (here of size 100) used to build the initial guess for B. Line 1 corresponds to a usual implementation for Davidson-like as schemes, with step 1B seldom used: the diagonal operator (eq 3) appears robust (small number of cycles) and faster than others. Line 2 is the choice stressing speed, with more frequent use of step 1B: expenses are reduced as expected but decrease in robustness (more cycles) except for the Gauss-Seidel triangular operator (Equation 4). Line 3 is the choice maximizing robustness, no matter the cost: now the triangular operator reveals its intrinsic stability and outperforms the diagonal operator. Line 4 is our final choice, dictated for automated usage, with definite preference for robustness under the constraint of a reasonable cost.

Table 2. Relative Cost Ratios Referenced to the Direct Diagonalization Method (QR) of a Matrix of Order $M = 100$[a]

		size *M*				
QR		100	200	400	800	1200
		1.0	7.7	58.3	432.3	1924.5
l	1	0.1	0.2	0.8	3.0	6.6
	5	0.6	1.7	4.8	16.0	38.2
	10	2.1	5.5	11.8	43.9	95.4
	20	7.8	20.7	37.8	120.1	258.6
	40	14.8	43.7	110.8	352.8	609.9

[a] The number of eigenvectors calculated is *l*.

Table 2 lists the performances of our implementation as a function of *M* and *l*. The direct diagonalization (QR) increases roughly as M^3, while the indirect method increases asymptotically as $l * M^2$. However, better performance is observed for small values of *l*, a consequence of the strategy of a simultaneous search for the target set of *l* eigenstates.

To conclude this section, the two following general principles can be drawn. (Note that the final implementation in AMPAC accounts for these points automatically during a calculation.)

1. The indirect algorithm is faster for *l* less than 10% of *M*.

2. The indirect algorithm becomes the only practical route for *M* greater than 500.

III. SORT AND SELECTION OF DETERMINANTS

For numerous applications of semiempirical methods usually focusing on ground state properties and most transition states, there is no need for a large CI active space. Thus, many important calculations can be completed using the indirect diagonalization described in the previous section, accounting for the full list of determinants that are generated. The effective limit for this procedure is reached near an active space of size 8 for CAS-CI and 15 for CISD or about 5000 determinants. Beyond this limit (easily exceeded in spectroscopic studies of large molecules and some other cases) a truncation of the full list of determinants must occur to control the combinatorial explosion if the calculation is to be feasible.

The selected determinants must provide a proper representation of the spin operator S^2. Each eigenstate must also be an eigenvector of both the Hamiltonian (**A**) and the representation of S^2. This property must hold not only for **A** but also for its ULC, the eigenvectors of which define the first model space **B** for the indirect method of diagonalization as described in the previous section. Practically, the deter-

minants must be collected as nonseparable classes, where each class introduces at least one complete set of determinants differing only by a spin inversion.

For each determinant $|i\rangle$, two energies may be defined: (1) the "true" energy $\langle i|\mathbf{A}|i\rangle$ hereafter named Epstein−Nesbet (EN) energy and (2) the Moller−Plesset (MP) energy $\langle i|\mathbf{F}|i\rangle$ where **F** is the Fock operator at SCF convergence. A useful property of the MP expression is that determinants differing only by spin or those sharing degenerate MOs (belonging to an irreducible representation with more than one member (i.e. **E** or **T**)) have the same energy. The MP energies therefore provide a simple way to collect determinants into classes. Thus, once a determinant is selected by some criterion, then its entire class is immediately appended to the description.

During the generation of a CAS description, the full list of determinants results from a direct product between the permutations of α spins and those of β spins. (A singles/doubles (CISD) generation follows analogous rules.) Another property of each MP energy is that it is equal to the sum of the contributions from the α spins and the β spins, with none of the coupling terms present in the parent EN energy. The full list of determinants rapidly becomes far too large to work with effectively as a whole past certain limits (active space of size larger than 9 for CAS-CI, 25 for CISD). In this case, the list of determinants used must be shortened in some reasonable fashion for computational feasibility. The determinants are collected by the MP energies into general classifications, without the initial requirement of an expensive sorting procedure within each of these classifications. Only determinants with higher MP energies are exactly sorted, making use of a "quick sort" algorithm working separately on the α and β MP energies. The resulting "short" list, which is defined to be 11 times the size of the final list M, is the one where a more refined and restrictive criterion of selection can be successfully applied. Thus, this procedure selects and retains only those determinants that have a reasonable chance of being included in and important to the final set.

The most important criterion for selection of a particular determinant is that determinant's EN energy; the ULC elements of **A** (of size *m*) are constructed based on this. Another option to consider is whether to retain these same criteria to direct completion of the matrix **A** up to the allowed size *M*. (This direction is the choice implemented on older versions of the semiempirical programs.[13]) The alternative to this is a multireference perturbative approach that may be of the MP[25] or the EN[26] type. The two approaches differ in the definition of the zero-order Hamiltonian. In the multireference approach, the Fock operator for MP and EN is defined by its representation, that is the diagonal part of the matrix **A**.

In a perturbative approach to the CI problem, the MP perturbative series is generally preferred because of its "size-consistency".[27−30] However, this property is in the context of poor performance with respect to the greater accuracy of second-order EN energy corrections, since a variational treatment of CI is not formally size-consistent. In the first implementation of a selection by perturbation,[31] we therefore preferred and implemented the EN approach, taking as criterion for the selection of a determinant the condition that $|k\rangle$ did not belong to the ULC of **A**, the L1 norm[32] of the first-order correction on vectors with respect to a "germ" of the *n* lowest determinants $|i\rangle$ belonging to the ULC:

$$W(k) = \left|\sum_{i=1}^{n} \frac{\langle i|A|k\rangle}{(\langle i|A|i\rangle - \langle k|A|k\rangle)}\right| \qquad (5)$$

The determinants $|k\rangle$ are sorted in decreasing order of the criterion $W(k)$ and are appended by classes to the ULC, from $m + 1$ up to the prescribed limit M for the matrix **A**.

Using relatively small values for *m* and *n* (typically 11 and 5, respectively), the selection is deliberately oriented toward the description of the ground state or the first few excited states with high components similar to the germ. It actually yields a good description for the (ground) transition states,[33] systems with numerous open shells[34] and complex (contributions from many determinants) first excited states.[35] But, this choice is poor for the calculation of UV/visible transition energies because the germ is no longer representative of the target, i.e. the *l* lowest exact eigenstates. Attempts to refine the values of *m* and *n* result in significant improvements in the prediction of these properties. However, there is certainly no clear predictive rule for the correct general values of *m* and *n*, aside from the fact that *m* must certainly be increased. Results of changes in *n* are irregular.

A very flexible and powerful approach to the variational treatment of a set of transition energies is the CIPSI algorithm,[36] which ends with an indirect diagonalization of a large CI matrix.[37] Besides its iterative aspect, the main difference here with respect to the strategy in the previous implementation in AMPAC 5.0[31] is the replacement of a germ space of *n* determinants with the set of the *n* first eigenvectors $|b\rangle$ of the ULC.

We tested CIPSI (beginning with the construction of the one- and two-electron semiempirical integrals) on a sample of six molecules (**A**−**F** in the Supporting Information) where the UV−visible spectra were poorly predicted using the previous approach.[31] The new method provides significant improvements and substantially dampens the chaotic behavior. Moreover, with an active space of roughly 15 MOs, a target space of *l* near 15, and a final CI matrix of size *M* approximately 1000, only one cycle of the CIPSI algorithm provides good results. This is without the need for subsequent cycles as soon as the starting ULC becomes of size *m* significantly larger than *l*.

We therefore implemented one CIPSI cycle in the procedure to select determinants, retaining the EN criterion for selection. In addition, we checked (for the sample of six molecules) the conditions generated by six criteria:

- **L1**, **L2**, **L**∞ norms on the first-order correction on the vectors
- **L1**, **L2**, **L**∞ norms on the second-order correction on the energies

The best results are somewhat correlated with the L1 norm energies as selected by our algorithm and PERTCI (which uses energy of the determinant only as a selection criterion).[10] The final criterion vector **W** associated with the determinants $|k\rangle$ therefore becomes

$$W(k) = \sum_{b=1}^{n} \frac{\langle b|A|k\rangle^2}{(E_b - (\langle k|A|k\rangle))|} \qquad (6)$$

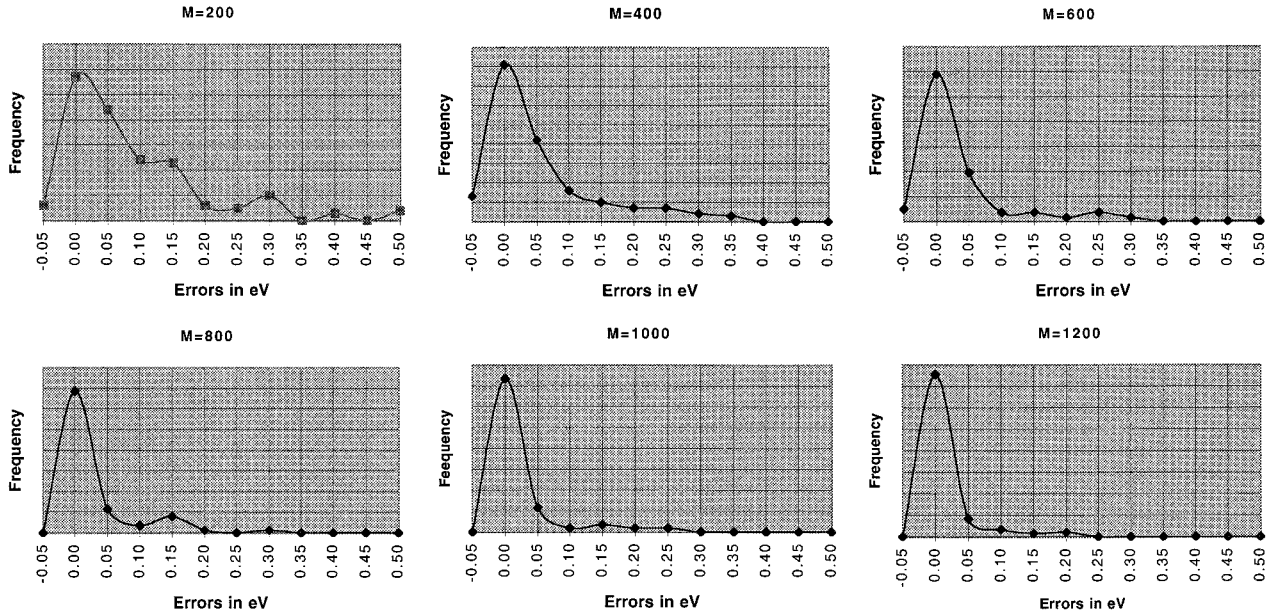

Figure 1. Error distributions of transitions of truncated CI at various levels vs full CI. (Conditions: 19 molecules, $l = 10$, $m = 150$, $n = 30$.)

The complete strategy for generation and selection of M "important" determinants thus finally depends on two parameters: m and n. The procedure is as follows.

Step 1. Generate by classes the short list by increasing sort of MP energies up to size L (where L is 10−12 times larger than M).

Step 2. Select by classes the ULC (size m) by increasing sort of EN energies.

Step 3. Diagonalize directly the ULC, producing eigenvectors $|b\rangle$ and energies E_b.

Step 4. For k from $m + 1$ to L, set the criterion vector **W** (Equation 6) with b from 1 to n.

Step 5. Generate by classes the final list (size M) by decreasing sort of **W**.

Each sort manipulation is best carried out by a variant of the quick-sort algorithm[15] adapted to this case of a selection by classes. As applied here, the most time-consuming part of the procedure is Step 4, calculation of numerous CI matrix elements $\langle i|\mathbf{A}|k\rangle$. The moderate expense at Step 3 is largely compensated for by the subsequent rapid and stable convergence of the indirect diagonalization strategy, since the eigenvectors $|b\rangle$ provide a well-behaved trial guess for the matrix **B**.

IV. CALIBRATION OF THE METHOD

The only remaining undefined part of the method is the selection values for m and n. These values could, in turn, be dependent on M and l. In spectroscopic calculations, the transitions of interest are usually those with energies of less than 5 eV. Alternatively, investigators are also interested in the eigenstates up to the first moderately intense UV band to model photochemical processes. For most unsaturated organic molecules, this suggests a typical value of $l = 10$ which is accepted as a standard. This corresponds to the 10 lowest transition energies. The AM1, MNDO, PM3, and SAM1 semiempirical models were used for these computations.

The definition of some statistical criterion or index of "quality" is required for benchmarking various trials prior to attempting to calibrate m and n (with M as a parameter). It must be noted in the present case that the criterion we are seeking is not (as of yet) the reliability of a particular semiempirical parametrization with respect to spectroscopic experimental data but the confidence in the *truncation* of the full list of the determinants. The only way to do this is to define an error vector by direct comparison of the "exact" transition energies (*no truncation* in the CI matrix) with those obtained in the presence of a truncation. The size of the sample test was extended to 19 (**A**−**S** in the *Supporting Information*) molecules, deliberately including transition states and molecules with open shells in their ground states. (No improvement in the results was obtained with a larger test set.) The active space (eight active MOs with CAS-CI, 15 MOs with CISD) yields an exact CI matrix of size about 5000.

Regardless of the values for m and n within large ranges, the distribution of errors always shows the same trend, as pictured in Figure 1. Most of the results are good with error near zero. A few are underestimated (negative error); this occurs if and only if the truncation is severe (M lower than 800). Most of the nonzero errors are positive. The mean error appears to be statistically independent of m and n, but strongly dependent on l, with a correlation coefficient = 0.9955. This bias toward overestimated transition energies is not a systematic failure of the method, but a consequence of MacDonalds theorem,[38] where eigenvalues of the truncated CI matrix are upper bound for the exact eigenvalues. Practically, the truncation makes the highest eigenvalues poorer than the lowest ones so that the mean signed error becomes positive.

Accounting for this bias, the index of quality that best fits our purpose is the root-mean-square (RMS) error. The RMS error was calculated systematically as a function of m, n, and M. This is a "brute force", but effective, approach to the minimization of a relatively noisy function. Two typical

Table 3. Two Cross-Sections ($M = 400$, $M = 1200$) of the RMS Error (eV) as a Function of m and n[a]

		n				
		10	15	20	30	50
		$M = 400$				
m	15	0.221	0.162			
	30	0.164	0.137	0.129	0.091	
	50	0.160	0.122	0.123	0.092	0.114
	100	0.158	0.120	0.124	0.099	0.113
	150	0.148	0.120	0.117	0.101	0.113
	all	0.162	0.162	0.162	0.162	0.162
		$M = 1200$				
	15	0.069	0.099			
	30	0.051	0.056	0.045	0.038	
	50	0.051	0.046	0.045	0.038	0.040
	100	0.051	0.046	0.044	0.037	0.040
	150	0.051	0.045	0.044	0.037	0.038
	all	0.043	0.043	0.043	0.043	0.043

[a] A sample of 19 molecules was used for 10 transitions in a CAS-CI with an active space of size 8.

cross-sections are reported in Table 3. From these results, it appears that a plot of the data is flat and a minimum near $n = 30/m = 100$ exists, and is insensitive to the value of M. The location of the minimum is also insensitive to the size l of the target for l in the range between 6 and 13. At low values of l, the minimum migrates irregularly toward $n = 15/m = 150$. Another conclusion to be drawn from these results is that a selection of the determinants solely based on the EN energies must be discarded given the quality of the CIPSI approach, especially for low values of M.

Similar trends and optimum values for m and n are noted when a CISD is performed. As M increases, the RMS error (as above, in comparison to the exact CISD limit) decreases more rapidly than the parent CAS-CI. This is not necessarily unexpected, as the perturbative criterion (Equation 6) scrutinizes a list of determinants $|k\rangle$ always in direct interaction with the ULC.

In the interest of a robust and reliable implementation of the indirect algorithm method, and in view of the preceding optimum results, we recommend the choice $n = 30/m = 100$) for studies with values for l from 1 (only the ground state) up to 13. This is a sufficient range for photochemistry and ordinary spectroscopic investigations. These values are implemented as the defaults in AMPAC 6.0.[14]

In certain cases, larger values for l could require larger values for m and n: In such a case, m is increased by a factor dependent on the problem, and n set to triple l. These defaults can be overridden by the user using AMPAC's PERTU (n,m) keyword. Extreme care must be exercised at this point.

V. CONCLUSION

We here propose an approach to variational CISD or CAS-CI problems in a semiempirical context based on a selection of determinants using the CIPSI algorithm and a pseudo-Davidson indirect diagonalization method. The basic objective was the development of a tool suitable for spectroscopic studies in general semiempirical methods. However, the strategy is also optimized for robustness in automated uses, such as intensive searches on potential energy surfaces for either ground or excited states.

Table 4. RMS Error (eV) as a Function of l and M with $n = 30/m = 100$[a]

		M					
		200	400	600	800	1000	1200
l	4	0.131	0.085	0.066	0.050	0.035	0.026
	7	0.134	0.088	0.067	0.052	0.038	0.029
	10	0.154	0.099	0.077	0.061	0.048	0.037
	%<2%	60%	82%	86%	88%	90%	94%

[a] The last line gives the percentage of relative errors lower than 2% ($l = 10$).

The intrinsic accuracy of the method is summarized in Table 4, which shows the RMS error as a function of the number of states calculated l and the size M of the CI matrix involved. The data indicates a roughly hyperbolic decay in M, ending below 0.04 eV with $M = 1200$. At this point, less than 2% of the calculated transition energies deviate statistically from their exact values by more than 0.15 eV.

These trends and confidence limits are best expressed by examining and comparing the worse errors in percent of UV−visible wavelengths. The worst relative error is 20% ($M = 200$), 10% ($M = 400$, the default value in the new implementation[14]), and 4% ($M = 1200$). While this relationship is linear, it cannot be invoked to "correct" empirically the wavelengths by a scaling factor, since it is derived from only a small selection of the set of wavelengths. Most transition energies are reproduced much more accurately that the maximum errors noted above. At the threshold of an error not exceeding 2% they increase in number from 60% at $M = 200$ to 94% at $M = 1200$ as shown on the last line of Table 4.

The performance of CIS and CISD approximations to a CAS-CI may also be quantified from the same set of molecule and active space calculations over 10 transitions. (No truncation in the CI matrix occurred in this comparison.) The results are as follows.

CIS. CIS is a very fast method but completely fails in the description of any excited states with a doubly excited dominant configuration. As these are often present in the first 10 excited states, the RMS error with respect to the CAS limit is quite large at 0.49 eV and a worst error of 2.97 eV.

CISD. With CISD, the RMS error falls to 0.30 eV, with the worst error at 0.71 eV. This shows clearly that CISD is an approximation to CAS-CI that is much poorer than the truncation procedure described here. Therefore, CISD is not recommended for spectroscopic studies.

However, CISD *is* useful for describing reaction paths and transition states on the ground-state surfaces of "normal" systems with no open shell electronic structures that lead to situations where a more complete description is required.[39] The important aspect of this from a practical implementation and utility perspective is that as truncation of the CI matrix occurs, much less "noise" is introduced in the calculation of the energy than was the case in the previous implementation in earlier versions of AMPAC. This continuous and smooth behavior is of significance for accurate and rapid convergence of various geometry optimizers on ill-behaved PESs.

The total cost (SCF plus CAS-CI) of the method is given in Table 5 as a function of M and l. The test case is a molecule with 35 atoms and an active space of 15 molecular

Table 5. Total Cost Ratios Referred to a CAS-CI Truncated at Size $M = 100$ and Diagonalized Directly[a]

		M				
		100	200	400	800	1200
l	1	0.96	1.01	1.70	2.06	2.52
	5	0.97	1.06	1.90	2.60	3.80
	10	1.00	1.22	2.18	3.78	6.15
	20	1.00	1.28	3.25	6.88	12.84
	40	1.00	1.28	4.09	16.44	27.25

orbitals. The cost increases slowly with both parameters, culminating at a ratio of 27.25 for $l = 40$ and $M = 1200$. Such a prohibitively expensive ratio is representative of a UV/visible calculation carried out for extreme accuracy and usually only performed in "single point" mode. In this case, the cost increase is actually negligible.

However, the real expense in semiempirical calculations involves optimization of geometries, location of transition states, etc. These are almost invariably conducted with reasonable values of *M* and at ground electronic states, i.e. where $l = 1$. With a ratio not exceeding 2.6 for $n = 1000$, the increase in timing is not significant. Moreover, an increase in *M* results in no change in the time spent in the analytical calculation of the CI molecular gradient. This is largely due to the implementation of the **Z**-vector technique and one- and two-particle density matrix formalism[40] in release 6.0 of AMPAC.[14] The performance of the method is similar to those reported elsewhere.[41]

ACKNOWLEDGMENT

We would like to thank all of the investigators who so kindly helped us by using the program and reporting to us the weaknesses of the selection by perturbation approach for large spectroscopic studies in release 5.0 of AMPAC.

Supporting Information Available: AMPAC/MOPAC style input files for the 19 molecules used to test methodology. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bingham, R. C.; Dewar, M. J. S.; Lo, D. H. *J. Am. Chem. Soc.* **1975**, *97*, 1285, 1294, 1302, 1307.
(2) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
(3) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209, 221.
(4) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(5) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391.
(6) Dewar, M. J. S.; Jie, C.; Yu, J. *Tetrahedron* **1993**, *49*, 5003.
(7) Thiel, W. *J. Am. Chem. Soc.* **1981**, *103*, 1413.
(8) Del Bene, J.; Jaffe, H. H. *J. Chem. Phys.* **1968**, *48*, 1807. Del Bene, J.; Jaffe, H. H. *J. Chem. Phys.* **1969**, *50*, 1126. Del Bene, J.; Jaffe, H. H. *J. Chem. Phys.* **1968**, *49*, 122. Del Bene, J.; Jaffe, H. H. *J. Chem. Phys.* **1968**, *48*, 4050.
(9) Ridley, J.; Zerner, M. C. *Theor. Chim. Acta (Ber.)* **1973**, *32*, 111.
(10) (a) Hase, H. L.; Lauer, G.; Schuete, K. S.; Schweig, A. *Theor. Chim. Acta* **1978**, *48*, 47. (b) Lauer, G.; Schuete, K. W.; Schweig, A. *J. Am. Chem. Soc.* **1978**, *100*, 4825. (c) Lauer, G.; Schuete, K. W.; Schweig, A.; Thiel, W. *Theor. Chim. Acta* **1979**, *52*, 312. (d) Schweig, A.; Thiel, W. *J. Am. Chem. Soc.* **1981**, *103*, 1425.
(11) (a) Berthier, G. *Compt. Rend. Acad. Sci.* **1954**, *238*, 91. *J. Chim. Phys.* **1954**, *51*, 363. (b) Pople, J. A.; Nesbet, R. K. *J. Chem. Phys.* **1954**, *22*, 571.
(12) (a) Longuett-Higgins, H. C.; Pople, J. A. *Proc. Phys. Soc.* **1955**, *A68*, 591. (b) Dewar, M. J. S.; Hasmall, J. A.; Venier, C. G. *J. Am. Chem. Soc.* **1968**, *90*, 1953.
(13) Stewart, J. J. P. MOPAC 7 (QCPE 455 7.0).
(14) AMPAC 6.0, Semichem, 7204 Mullen, Shawnee, KS 66216, 1997 .
(15) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling; W. T. *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: 1986.
(16) Lanczos, C. *J. Res. Nat. Bur. Stand.* **1950**, *45*, 255.
(17) (a) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87. (b) Davidson, E. R. *Comput. Phys. Commun.* **1989**, *53*, 49.
(18) Gadea, F. X. *Chem. Phys. Lett.* **1994**, *227*, 201.
(19) Ortega, J. *Mathematical Methods for Digital Computers*, John Wiley: 1960.
(20) Weltin, E. *Int. J. Quantum Chem.* **1969**, *3*, 635.
(21) Fadeev, D. K.; Fadeeva, V. N. *Computational Methods of Linear Algebra*; W. H. Freeman & Co.: San Francisco, 1963.
(22) Dewar, M. J. S.; Liotard, D. *J. Mol. Struct. (THEOCHEM)* **1990**, *206*, 123.
(23) Patchkovskii, S.; Thiel, W. *J. Comput. Chem.* **1996**, *17*, 1318.
(24) Olsen, J., Joregensen, P.; Simons, J. *Chem. Phys. Lett.* **1990**, *169*, 463.
(25) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
(26) (a) Epstein, P. S. *Phys. Rev.* **1924**, *28*, 695. (b) Nesbet, R. K. *Proc. R. Soc.* **1955**, *A230*, 312; *Proc. R. Soc.* **1955**, *A230*, 922.
(27) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 281.
(28) Pople, J. A.; Seeger, R.; Krishnan, R. *Int. J. Quantum Chem. Symp.* **1977**, *11*, 149.
(29) Krishnan, R.; Pople, J. A. *Int. J. Quantum Chem.* **1978**, *14*, 91.
(30) Krishnan, R.; Frisch, M. J.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 4244.
(31) AMPAC 5.0; Semichem, 7204 Mullen, Shawnee, KS 66216, 1994.
(32) The L1 norm of a vector is the sum of the absolute values of its components; the L2 norm is the usual Euclidean norm, while the L∞ norm is the largest component in absolute value.
(33) Pons, J. M.; Oblin, M.; Pommier, A.; Rajzmann, M.; Liotard, D. *J. Am. Chem. Soc.* **1997**, *119*, 3333.
(34) Dannenberg, J. J.; Liotard, D.; Halvick, P.; Rayez, J. C. *J. Phys. Chem.* **1996**, *100*, 9631.
(35) Marchand, N.; Jimeno, P.; Rayez, J. C.; Liotard, D. *J. Phys. Chem.* **1997**, *101*, 6077.
(36) Huron, B.; Malrieu, J. P.; Rancurel, P. *J. Chem. Phys.* **1973**, *58*, 5745. CIPSI stands for Configuration Interaction by Perturbation with multiconfigurational zeroth-order wave function Selected by Iterative process.
(37) Evangelisti, S.; Daudey, J. P.; Malrieu, J. P. *Chem. Phys.* **1983**, *75*, 91.
(38) (a) Hyleraas, E. A.; Undheim, B. *Z. Phys.* **1930**, *65*, 759. (b) Mac Donald, J. K. L. *Phys. Rev.* **1933**, *43*, 830.
(39) Dannenberg, J. J.; Liotard, D. Work in progress.
(40) Handy, N. C.; Schaeffer, H. F., III *J. Chem. Phys.* **1984**, *81*, 5031.
(41) S. Patchkovskii, W. Thiel, *Theor. Chim. Acta* **1996**, *93*, 87.
(42) [a] *l* is the number of calculated states.

CI980149I