Multi-Coefficient Correlation Method for Quantum Chemistry

Patton L. Fast, José C. Corchado, María L. Sánchez, and Donald G. Truhlar*

*Department of Chemistry and Supercomputer Institute, University of Minnesota,
Minneapolis, Minnesota 55455-0431*

Received: January 29, 1999

We present a new method for extrapolating correlated electronic structure calculations based on correlation-consistent polarized double-ζ and triple-ζ basis sets for calculation of molecular energies (atomization energies).

1. Introduction

Advances in computational sciences as well as the development of new algorithms are making it possible to carry out ab initio electronic structure calculations on small systems with errors approaching the accuracy of experimental measurements and with even better accuracy than experiment in a small but rapidly growing number of cases. Nevertheless, we are still far from being able to make reliable quantitative predictions based on ab initio calculations for large and even medium-sized systems. Even though in theory the algorithms are applicable to any system despite its size, the computational resources required for carrying out such calculations for medium- and large-size systems are beyond the scope of available technology, and for large systems we can assume that this situation will continue for a long time. Thus, it is necessary to develop methods that can be applied to medium- and large-sized systems with a reduced computational cost.

The two major sources of error in an ab initio calculation of molecular energies are the truncation of the one-electron basis set and the truncation of the number of excitations or configurations used for treating correlation energies. In principle, one could calculate the energy of a system of interest using a relatively small basis set and including a reduced number of configurations or excitation operators in the calculation of the correlation energy, and then improve both the basis set and the configuration or excitation space until convergence is achieved. This convergence could be measured as the difference between the approximate calculation and the exact solution to the Schrödinger equation, namely a full configuration interaction (FCI) calculation using an infinite basis set. This combination is called complete configuration interaction (CCI). However, for most systems of interest, the computational cost of either FCI or CCI makes them impossible. One promising way to circumvent this difficulty is to calculate the first few terms in a sequence of improving calculations and use these data to extrapolate to the CCI limit. In the present paper we propose a new set of semiempirical methods designed to do this as accurately as possible. Four recent reviews may be consulted for a summary of available methods for extrapolation.[3−6] The developments reported in the present paper were motivated by two of the previous extrapolation methods, namely the scaling all correlation (SAC) method[5−9] introduced by Gordon and one of the authors and the ab initio infinite basis (IB) method[l0,11] discussed recently. We note that the SAC method is itself based on the earlier scaling external correlation (SEC) method[6,12] of Brown and one of the authors, and the IB method is based on earlier work[4,13−15] extrapolating correlation-consistent basis sets. In fact, the systematic convergence[4,13−17] of correlation-consistent basis sets[18,19] is believed to be a key ingredient in the success of the method proposed here.

Section 2 presents some useful notation. Section 3 uses this notation to describe all methods considered in this paper, including the new extrapolation method, which we call the multi-coefficient correlation method (MCCM).

2. Notation

Throughout this paper we will use the pipe "|" to represent the energy difference either between two one-electron basis sets B1 and B2 or between two many-body levels L1 and L2, e.g., Møller−Plesset second-order perturbation theory and Hartree−Fock theory. The energy difference between two basis sets will be represented as

$$\Delta E(\mathrm{L/B2|B1}) \equiv E(\mathrm{L/B2}) - E(\mathrm{L/B1}) \qquad (1)$$

where L is a particular electronic structure method and B1 is smaller than B2. The energy change that occurs upon increasing the treatment of the correlation energy will be represented by

$$\Delta E(\mathrm{L2|L1/B}) \equiv E(\mathrm{L2/B}) - E(\mathrm{L1/B}) \qquad (2)$$

where L1 is a lower level of theory than L2 and B is a common basis set. Finally, the change in energy increment due to increasing the level of the treatment of the correlation energy with one basis set as compared to the increment obtained with a smaller basis set will be represented as

$$\Delta E(\mathrm{L2|L1/B2|B1}) \equiv E(\mathrm{L2/B2}) - E(\mathrm{L1/B2}) - [E(\mathrm{L2/B1}) - E(\mathrm{L1/B1})] \qquad (3)$$

All new calculations in this paper are based on three correlation-consistent basis sets,[18,19] namely cc-pVDZ, aug″-cc-pVDZ (we use the double prime notation to denote that diffuse functions have been omitted on hydrogen and that the diffuse subshell corresponding to the highest angular momentum has been omitted for the heavy atoms, i.e., omitting the diffuse *d* function from the aug′-cc-pVDZ[20] basis set), and cc-pVTZ. Since we restrict ourselves to only three basis sets, no confusion can result from a shorthand notation, and we call these pDZ, pDZ+, and pTZ, respectively. In addition, in some cases we will compare to previous calculations based on Pople-type basis sets such as 6-311G** which are explained elsewhere.[21]

The new methods, explained in section 3, will have names like MCSAC-L, MCCM-L, and MCCM-L2;L1. The energies

10.1021/jp9903460 CCC: $18.00
Published on Web 06/15/1999

in these methods consist of sums of terms, most of which have the form of eq 1, 2, or 3.

We will use standard abbreviations for electronic structure methods. All new calculations in this paper are based on the following methods:In addition, in some cases we will compare to previous calculations based on the QCISD(T)[24] method. This denotes quadratic configuration interaction with single and double excitations plus two quasiperturbative terms involving triple excitations. Thus, in formulas, L, L1, L2, and L3 denote one or another of HF, MP2, MP4D, MP4SDQ, and so forth.

HF	Hartree−Fock[21]
MP2	Møller−Plesset (MP) perturbation theory, second order[21]
MP4D	MP perturbation theory, fourth order, with double excitations[21]
MP4SDQ	MP perturbation theory, fourth order, with single, double, and quadruple excitations[21]
MP4	full fourth-order MP perturbation theory, i.e., MP4SDQ plus triple excitations[21]
CCSD	coupled-cluster theory with single and double excitations[22]
CCSD(T)	CCSD plus two quasiperturbative terms involving triple excitations[23]

A critical element in what follows will be the idea of a *sequence* of correlated calculations. We will define two sequences: the MP sequence and the CC sequence. The MP sequence consists of $L_0 = HF$, $L_1 = MP2$, $L_2 = MP4D$, $L_3 = MP4SDQ$, and $L_4 = MP4$; and the CC sequence consists of $L_0 = HF$, $L_1 = MP2$, $L_2 = CCSD$, and $L_3 = CCSD(T)$. Note that if one performs a calculation of the energy by any of the methods in a sequence, the energy for each of the lower levels in the sequence is also available for no additional cost. Thus, for example, if one asks the *Gaussian94* program[25] to carry out a CCSD calculation, it also writes the HF and MP2 energies since these are calculated as intermediate steps in a CCSD calculation.

All ab initio calculations considered in this paper are frozen-core methods. When one uses such methods, FCI means frozen-core, full-valence configuration interaction, and CCI means frozen-core, complete-valence configuration interaction. The assumption implicit in using frozen-core calculations is that core energies cancel out in calculating bond energies, atomization energies, barrier heights, and other potential energy surface features. This is not perfectly correct;[26] therefore, we will add in the core-correlation energy by the method described by two of the current authors which is described elsewhere.[27] (This method is parametrized for molecules containing H, Li, Be, B, C, N, O, F, Al, Si, P, S, and Cl, and includes both core−core and core−valence correlation contributions.)

3. Methods

Section 3.1 reviews the SAC[5−9] and ab initio IB[10,11] methods in terms of the notation introduced above, as motivation for the new methods. It also introduces the new empirical infinite-basis method (EIB). Section 3.2 presents the new set of MCSAC and MCCM methods. Section 3.3 summarizes the Gaussian-1 and Gaussian-2 (G1 and G2) methods in terms of the notation of section 2, for comparison purposes.

3.1. SAC, IB, and EIB. The SAC method may be written

$$E(\text{SAC-L/B}) = E(\text{HF/B}) + c_1 \Delta E(\text{L}|\text{HF/B}) + E_{\text{SO}} + E_{\text{CC}} \quad (4)$$

where, in the original notation,[6−9] the coefficient c_1 was written as $1/F$, where F is a parameter. The value of F, or equivalently c_1, is determined by comparison to experimental data.[6−9] In the present formulation of the SAC method we explicitly include the spin−orbit and core-correlation contributions to the energy, E_{SO} and E_{CC}. The formulation of the SAC method given in ref 9 includes the E_{SO} term, but not the E_{CC} term, and even earlier[6−8] applications neglect both E_{SO} and E_{CC}. Therefore, caution should be taken when making a direct comparison between the formulation and results given here and those given in earlier papers.

The ab initio IB method may be written

$$E(\text{IB-L/B2}|\text{B1}) = E(\text{HF/B1}) + c_1 \Delta E(\text{HF/B2}|\text{B1}) + \Delta E(\text{L}|\text{HF/B1}) + c_2 \Delta E(\text{L}|\text{HF/B2}|\text{B1}) + E_{\text{SO}} + E_{\text{CC}} \quad (5)$$

where, in the original notation,[10] with B1 = pDZ and B2 = pTZ, c_1 is given by $3^{\alpha}/(3^{\alpha}-2^{\beta})$ and c_2 is given by $3^{\beta}/(3^{\beta}-2^{\beta})$, where α and β are parameters. The values of α and β, or equivalently c_1 and c_2, were determined[10] by comparison of the first four terms of eq 5 to ab initio estimates of the infinite-basis-set limit of correlation method L.

The empirical infinite-basis (EIB) method, introduced here, is defined to have the same form as eq 5, but now the coefficients are fit to experimental energy differences, such as atomization energies.

Notice that the SAC method and the EIB method, by virtue of being fit to experimental data, represent an attempt to extrapolate to the CCI limit, whereas the ab initio IB method attempts to reach the infinite-basis limit for a given electron correlation level.

3.2. MCSAC and MCCM. The overall methodology for SAC is to scale all of the correlation energy that comes from a given level of correlation energy treatment, but using a single basis. However, the different components of the correlation energy may need different scaling factors. For example, using MP4D theory and the pDZ basis set, we can write

$$E(\text{MCSAC-MP4D/pDZ}) = E(\text{HF/pDZ}) + c_1 \Delta E(\text{MP2}|\text{HF/pDZ}) + c_2 \Delta E(\text{MP4D}|\text{MP2/pDZ}) + E_{\text{SO}} + E_{\text{CC}} \quad (6)$$

where c_1 and c_2 are constants. In general, we define the multi-coefficient SAC method (MCSAC) for electron correlation level L_n and basis B as

$$E(\text{MCSAC-L}_n/\text{B}) = E(\text{HF/B}) + \sum_{m=1}^{n} c_m \Delta E(\text{L}_m|\text{L}_{m-1}/\text{B}) + E_{\text{SO}} + E_{\text{CC}} \quad (7)$$

where the zeroth level, L_0, is HF theory, and L_1, L_2, ... are the correlated members of the sequence leading up to level L_n. These sequences are defined in section 2. Implicit in both SAC and MCSAC is the assumption that the error in the unextrapolated calculation is primarily in the correlation energy (as opposed to the Hartree−Fock energy). This is clearly an assumption, but it is not unreasonable since the correlation energy converges more slowly than the HF energy as one increases the basis set.

Clearly one can combine the MCSAC and IB methods. Furthermore, we can put variable coefficients on all terms, even the first one. We will call this the multi-coefficient correlation

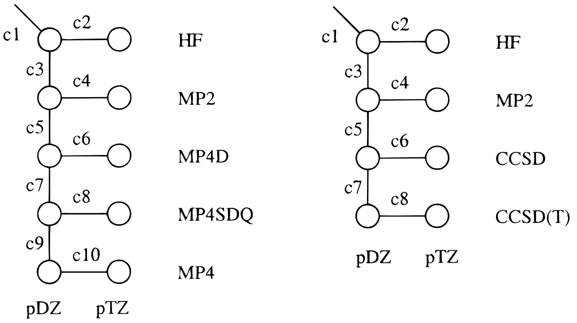

Figure 1. Coefficient trees for MCCM-MP4 and MCCM-CCSD(T). Coefficient trees for other symmetric MCCM methods (Colorado methods) are obtained by deleting rows successively from the bottom. (left) MP tree; (right) CC tree.

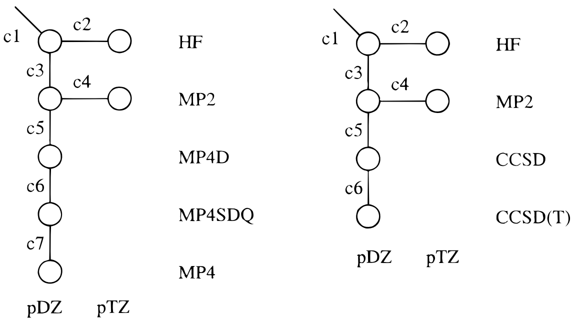

Figure 2. Coefficient tree for asymmetric MCCM-MP4;MP2 and MCCM-CCSD(T);MP2. Coefficient trees for other asymmetric MCCM methods (Utah methods) are obtained by deleting rows successively from the bottom. (left) MP tree. (right) CC tree.

method (MCCM). For example

$$E(\text{MCCM-CCSD}) = c_1E(\text{HF/pDZ}) + c_2\Delta E(\text{HF/pTZ|pDZ}) + c_3\Delta E\ (\text{MP2|HF/pDZ}) + c_4\Delta E(\text{MP2|HF/pTZ | pDZ}) + c_5\Delta E(\text{CCSD|MP2/pDZ}) + c_6\Delta E\ (\text{CCSD|MP2/pTZ|pDZ}) + E_{\text{SO}} + E_{\text{CC}} \tag{8}$$

where c_1 through c_6 are constants.

The coefficients and the corresponding energy differences can be visualized with Figure 1. The first circle represents E(HF/pDZ); the vertical lines represent level improvements (e.g., ΔE(MP2|HF/ pDZ)), and the horizontal lines represent basis set improvements (e.g., ΔE(MP2|HF/pTZ|pDZ)). Deleting the last row of the CC tree yields the tree corresponding to eq 8; comparing the CC tree to eq 8 should make the notation obvious. Equation 8 and Figure 1 involve rectangular arrays of circles; these can be called Colorado methods since Colorado is a perfect rectangle.

It is also useful to consider diagrams like Figure 2, which are rectangular with a corner missing, like Utah. For example,

$$E(\text{MCCM-CCSD; MP2}) = c_1E(\text{HF/pDZ}) + c_2\Delta E(\text{HF/pTZ|pDZ}) + c_3\Delta E\ (\text{MP2|HF/pDZ}) + c_4\Delta E(\text{MP2|HF/pTZ|pDZ}) + c_5\Delta E(\text{CCSD|MP2/pDZ}) + E_{\text{SO}} + E_{\text{CC}} \tag{9}$$

We hope the above examples have made the methods and the notation clear. We can summarize the methods more formally as follows:

$$E(\text{MCCM-L}_n) = c_1E(\text{HF/B1}) + c_2\Delta E(\text{HF/B2|B1}) + \sum_{m=1}^{n} c_{2m+2}\Delta E(\text{L}_m|\text{L}_{m-1}/\text{B2|B1}) + \sum_{m=1}^{n} c_{2m+1}\Delta E(\text{L}_m|\text{L}_{m-1}/\text{B1}) + E_{\text{SO}} + E_{\text{CC}} \tag{10}$$

and

$$E(\text{MCCM-L}_\upsilon;\text{L}_n) = c_1E(\text{HF/B1}) + c_2\Delta E(\text{HF/B2|B1}) + \sum_{m=1}^{n} c_{2m+2}\Delta E(\text{L}_m|\text{L}_{m-1}/\text{B2|B1}) + \sum_{m=1}^{n} c_{2m+1}\Delta E(\text{L}_m|\text{L}_{m-1}/\text{B1}) + \sum_{m=n+1}^{\nu} c_{m+n+2}\Delta E(\text{L}_m|\text{L}_{m-1}/\text{B1}) + E_{\text{SO}} + E_{\text{CC}} \tag{11}$$

By definition $\upsilon \geq n + 1$. Note that in this paper, B1 is always pDZ, and B2 is always pTZ.

We also consider adding a correction for diffuse functions:

$$E(\text{MCCM-L}_\upsilon;\text{L}_n;\text{HF+}) = E(\text{MCCM-L}_\upsilon;\text{L}_n) + c_{n+\upsilon+3}\Delta E(\text{HF/pDZ+|pDZ}) \tag{12}$$

This is a New Mexico method.

3.3. G1 and G2. In Gaussian-1 (G1) theory,[28] one directly sums the basis set and correlation energy increments without scaling, and then one adds two terms, together called the high-level correction, with empirical coefficients. The total G1 energy may be written as

$$E(\text{G1}) = E(\text{PDG1}) + c_1(n_\alpha - n_\beta) + \text{c}_2(n_\alpha + n_\beta) \tag{13}$$

where E(PDG1) is the properly dissociating contribution to the G1 energy,

$$E(\text{PDG1}) = E[\text{QCISD(T)/6-311G**}] + \Delta E(\text{MP4/6-311+G**|6-311G**}) + \Delta E[\text{MP4/6-311G**(2}df\text{)|6-311G**}] \tag{14}$$

and n_α and n_β are the number of α and β valence electrons (by definition $n_\alpha \geq n_\beta$). Note that the last term of eq 13 always cancels out in observable energy differences (such as bond energies), and so it is irrelevant for our purposes. Thus we consider the G1 method to be a one-parameter theory.

The term involving $n_\alpha - n_\beta$ in eq 13 is actually problematical. Consider, for example, the potential energy curve of C1$_2$. At the equilibrium internuclear distance, $n_\alpha - n_\beta = 0$, but for each separated atom, $n_\alpha - n_\beta = 1$. Thus, somewhere along the potential curve, as C1$_2$ is dissociated, $n_\alpha - n_\beta$ must change from 0 to 2. For this reason, G1 and G2 cannot be used to calculate globally continuous potential energy surfaces. The problem is similar, but not precisely the same as, violating size consistency; we call it the problem of improper dissociation. However, removing the $n_\alpha - n_\beta$ term greatly deteriorates the accuracy; this will be illustrated in a later section where we give mean errors for the properly dissociating Gaussian-2 (PDG2) and full G2 (denoted simply G2).

TABLE 1: Parameters Optimized in This Work

methods	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
SAC-MP2/pDZ	1.2768									
SAC-MP2/pTZ	1.0482									
SAC-MP4D/pDZ	1.4205									
SAC-MP4D/pTZ	1.1546									
SAC-MP4SDQ/pDZ	1.4189									
SAC-MP4SDQ/pTZ	1.1747									
SAC-MP4/pDZ	1.3243									
SAC-MP4/pTZ	1.0756									
SAC-QCISD/pDZ	1.4257									
SAC-CCD/pDZ	1.4690									
SAC-CCSD/pDZ	1.4375									
SAC-CCSD/pTZ	1.1915									
SAC-CCSD(T)/pDZ	1.3542									
SAC-CCSD(T)/pTZ	1.1100									
EIB-MP2	0.9172	1.5328								
EIB-MP4D	2.6023	1.1428								
EIB-MP4SDQ	2.4282	1.4396								
EIB-MP4	1.4401	1.3383								
EIB-CCSD	2.5931	1.4326								
EIB-CCSD(T)	1.8719	1.3256								
MCSAC-MP4D/pDZ	1.3942	1.1341								
MCSAC-MP4D/pTZ	1.1104	0.6460								
MCSAC-MP4SDQ/pDZ	1.4167	1.4076	2.1571							
MCSAC-MP4SDQ/pTZ	1.0579	1.5120	1.8980							
MCSAC-MP4/pDZ	1.3112	2.2512	2.3410	2.8643						
MCSAC-MP4/pTZ	1.0504	1.5680	1.8994	0.1483						
MCSAC-CCSD/pDZ	1.4174	1.2403								
MCSAC-CCSD/pTZ	1.1406	0.7433								
MCSAC-CCSD(T)/pDZ	1.3055	1.7800	3.0180							
MCSAC-CCSD(T)/pTZ	1.0513	1.2183	2.1835							
MCCM-MP2	0.9971	1.6560	0.7718	2.6398						
MCCM-MP4D	0.9893	1.7321	1.0640	1.2460	0.8177	0.1506				
MCCM-MP4SDQ	0.9642	1.6869	1.0099	2.4717	0.4514	3.4843	0.9629	3.5916		
MCCM-MP4	0.9893	1.4520	0.7228	2.5246	1.4028	4.7202	2.8677	4.8061	1.7804	7.8256
MCCM-CCSD	0.9703	1.6636	1.1206	1.7329	0.7127	2.6136				
MCCM-CCSD(T)	0.9887	1.5377	1.0048	1.5208	1.0106	1.5695	1.7202	0.9124		
MCCM-MP4D;MP2	0.9916	1.7056	1.0585	1.2314	0.7835					
MCCM-MP4SDQ;MP2	0.9862	1.5824	1.0618	1.4209	0.8955	1.1224				
MCCM-MP4;MP2	0.9803	1.1958	0.9559	1.9302	1.3053	1.7328	1.7792			
MCCM-CCSD;MP2	0.9949	1.6872	1.1422	0.8323	1.0040					
MCCM-CCSD(T);MP2	1.0002	1.4852	1.0026	1.1447	1.1869	2.1343				
MCCM-CCSD(T);MP2;HF+	1.0013	1.4127	0.3790	1.0173	1.0516	1.2159	2.1967			
G2′	2.2600									

In Gaussian-2 (G2) theory,[29] one adds additional terms to the G1 energy. The total G2 energy can be written as

$$E(\mathrm{G2}) = E(\mathrm{PDG2}) + c_1(n_\alpha - n_\beta) + \mathrm{c}_2(n_\alpha + n_\beta) \quad (15)$$

where E(PDG2) denotes the properly dissociating contribution to the total G2 energy,

$$\begin{aligned} E(\mathrm{PDG2}) = {} & E(\mathrm{PDG1}) + \\ & \Delta E[\mathrm{MP2/6\text{-}311{+}G}(3df,2p)|\mathrm{6\text{-}311G}(2df,p)] - \\ & \Delta E[\mathrm{MP2/6\text{-}311{+}G}(d,p)|\mathrm{6\text{-}311G}(d,p)] \quad (16) \end{aligned}$$

Again all results in this paper are independent of c_2.

In addition to G1 and G2, with the original parameters we consider a method we call G2′:

$$E(\mathrm{G2'}) = E(\mathrm{PDG2}) + c_1(n_\alpha - n_\beta) + E_{\mathrm{SO}} + E_{\mathrm{CC}} \quad (17)$$

Since eq 18 includes spin−orbit and core-correlation energies explicitly, we reoptimize c_1 for the G2′ method.

4. Determination of Parameters

All methods are tested in this paper using a 49-molecule data set presented previously.[9] In addition, the new methods are parametrized against this data set. The data set consists of all molecules in the original G2 data set that do not have any metal atoms. For each molecule the potential energy of atomization is obtained by combining the best experimental estimate of the heat of formation with an ab initio estimate of the vibrational energy released upon dissociation. The resulting atomization energies are the equilibrium dissociation energies D_e (this is the same notation as used in spectroscopy, but remember that in this paper D_e refers to complete dissociation of all bonds to form atoms). The D_e values are tabulated in ref 9, which also gives references for the experimental data and details of the vibrational energy calculations.

The structures of the 49 molecules were optimized at the MP2/pDZ level of theory. The optimized geometries were then used to calculate the HF and MP2 energies with the pDZ, pDZ+, and pTZ basis sets, the MP4 and CCSD(T) energies and their components (see section 2) with the pDZ and pTZ basis sets, and the QCISD and CCD energies with the pDZ basis set.

The spin−orbit term was calculated using experimental data as discussed in ref 9. The core-correlation energy was estimated by the very simple methods of ref 27. This method is so simple that it can be carried out on a hand-held calculator (or the back of an envelope), and it (like the E_{SO} estimate) contributes negligible cost. All coefficients for the new methods were then

TABLE 2: Mean Signed Errors, Mean Unsigned Errors, and rms Errors for the Nonparametrized Methods As Compared to the Accurate Dissociation Energies and the Corresponding Computational Cost for Each Method

method	MSE (kcal/mol)	MUE (kcal/mol)	RMSE (kcal/mol)	computational cost[a]
PDG1	−9.07	9.07	10.15	21
PDG2	−6.66	6.66	7.40	29
HF/pDZ	−90.10	90.10	99.08	0.5
HF/pDZ+	−89.57	89.57	98.65	0.5
HF/pTZ	−84.12	84.12	92.54	6.8
MP2/pDZ	−21.65	21.65	24.30	1.0
MP2/pDZ+	−20.62	20.62	23.26	0.9
MP2/pTZ	−5.40	6.74	8.53	14
MP4D/pDZ	−27.27	27.27	30.25	1.4
MP4D/pTZ	−11.95	11.95	14.05	20
MP4SDQ/pDZ	−27.22	27.22	30.04	1.4
MP4SDQ/pTZ	−12.96	12.96	14.77	20
MP4/pDZ	−22.91	22.91	25.19	2.0
MP4/pTZ	−6.45	6.45	7.47	42
QCISD/pDZ	−27.31	27.31	30.22	2.3
CCD/pDZ	−29.36	29.36	32.58	2.5
CCSD/pDZ	−27.84	27.84	30.85	3.1
CCSD/pTZ	−13.78	13.78	15.79	31
CCSD(T)/pDZ	−23.99	23.99	26.55	3.9
CCSD(T)/pTZ	−8.50	8.50	9.62	105

[a] Computational cost is the mean CPU time on an IBM SP computer for the six largest molecules, H_2CCH_2, H_3CCH_3, H_3COH, H_2NNH_2, Si_2H_6, and CH_3SH, relative to the mean CPU time (66 s) for an MP2/pDZ calculation on the same computer for the same six molecules.

optimized using linear regression against the data set of 49 accurate D_e values. The optimized coefficients are given in Table 1.

All electronic structure information, optimized geometries, electronic energies, and G2 energies for the 49 molecules and 9 atoms were obtained using the Gaussian94[25] electronic structure package.

5. Results

The mean signed error (MSE), mean unsigned error (MUE), and root-mean-square error (RMSE) for all the nonparametrized methods are given in Table 2. The nonparametrized methods include the Gaussian-1 and Gaussian-2 theories without the corresponding higher-level correction factors, denoted PDG1 and PDG2 to denote proper dissociation. The MSE, MUE, and RMSE for all the parametrized methods are given in Table 3. The computational cost in Tables 2 and 3 is the mean CPU time on an IBM SP computer for the six largest molecules, H_2CCH_2, H_3CCH_3, H_3COH, H_2NNH_2, Si_2H_6, and CH_3SH, relative to the mean CPU time for an MP2/pDZ calculation on these same six molecules. Computer times for smaller jobs may have significant components of overhead and should not be interpreted too closely.

6. Discussion

Comparing the multilevel, multi-basis-set PDG2 results with the single-level single-basis-set calculations shows that MP2/pTZ yields an MSE that is more than 1 kcal/mol better than PDG2 and an MUE that is slightly better than PDG2. The MP2/pTZ calculation has basically the same MSE and MUE as PDG2.

Comparing the MSE columns of Tables 2 and 3 for PDG1, PDG2, G1, and G2 we can see that adding the empirical corrections to PDG1 and PDG2 reduces the MSE by approximately 8.2 kcal/mol for PDG1 and approximately 6.6 kcal/mol for PDG2. In addition, the MUE is reduced by approximately 7.4 kcal/mol for PDG1 and approximately 5.4 kcal/mol for PDG2.

The MUE and RMSE for G2 and G2′ are essentially the same. The reoptimized value of c_1 is 2.26 and is very similar to the original value, 2.31; therefore, we can see that the high-level correction in G2 makes up for most of the spin-orbit and core-correlation effects.

Comparing the one-parameter SAC methods with their non-SAC counterparts shows that by simply scaling the correlation energy one can significantly reduce the errors. The MCSAC method further reduces the MUE and RMSE, but they require no further work. The EIB methods further improve the MUE and RMSE as compared to the MCSAC methods, but at the cost of running two basis sets. We note, however, that the cost of a second basis set is almost negligible. For example, the cost of calculating the six largest molecules with both pTZ and pDZ basis sets at the CCSD(T) level is only 4% higher than carrying out only the pTZ calculation. The relative cost increment is somewhat higher (5−10%) for other levels but still almost negligible.

The MCCM methods clearly outperform the corresponding SAC, MCSAC, and EIB methods. In fact, the MCCM-CCSD(T) method yields MSEs, MUEs, and RMSEs that are 0.1, 0.4, and 0.6 kcal/mol lower than G2. The major disadvantage to the MCCM-CCSD(T) method is the expense. However, by using the Utah strategy instead of the full Colorado strategy, we obtain the MCCM-CCSD(T);MP2 method for which the values of MSE, MUE, and RMSE are still slightly better than G2, but the cost is only 60% of the cost of the G2 calculation. We note that the MCCM methods achieve this better performance without the high-level correction, and thus they do not suffer from improper dissociation.

An important qualitative difference between Tables 2 and 3 is that the unparametrized methods show large negative mean signed errors in the range −5 to −90 kcal/mol in D_e. In contrast, the parametrized methods all have much smaller mean signed errors from +0.1 kcal/mol to −3 kcal/mol.

Table 4 shows the results obtained by the IB method of ref 10. This method gives much larger deviations from experiment than the new EIB methods. This is not surprising since the IB methods were not designed to yield experimental accuracy; they are designed to remove only the part of the error arising from an incomplete one-electron basis set.

An important aspect of Table 1 is that all coefficients are positive. We wrote the trees in such a way that positive coefficients correspond to physical fits. We did find that including certain combinations of methods, e.g., including both MP4DQ and MP4SDQ, sometimes gave unphysical fits, but all methods presented here have only positive coefficients. We attempted to include diffuse basis functions by using HF/pDZ+ and MP2/pDZ+ calculations. Including diffuse character at the MP2 level does not reduce the error and yields unphysical coefficients. However, adding diffuse character at the HF level helps to reduce the error and gives a physical fit when CCSD(T) energies are included.

The performance vs cost tradeoff of all the MCSAC, EIB, and MCCM methods is shown in Figure 3. The nine methods that lie below the line are our recommended methods, namely MCSAC-MP4SDQ/pDZ, MCSAC-MP4/pDZ, MCCM-MP4;MP2, MCCM-CCSD;MP2, MCCM-CCSD(T);MP2, MCCM-CCSD(T);MP2;HF+, MCCM-CCSD, MCCM-MP4, and MCCM-CCSD(T). The × in Figure 3 denotes both G2 and G2′.

TABLE 3: Mean Signed Errors, Mean Unsigned Errors, and rms Errors for the Parametrized Methods as Compared to the Accurate Dissociation Energies and the Corresponding Computational Cost for Each Method

method	parameters	MSE (kcal/mol)	MUE (kcal/mol)	RMSE (kcal/mol)	computational cost[a]
G1	1	−0.89	1.64	2.10	21
G2	1	−0.09	1.22	1.69	29
G2′	1	−0.24	1.21	1.68	29
SAC-MP2/pDZ	1	−2.70	9.47	11.64	1.0
SAC-MP2/pTZ	1	−1.61	5.88	7.40	14
SAC-MP4D/pDZ	1	−0.85	5.10	7.74	1.4
SAC-MP4D/pTZ	1	−0.80	4.30	6.68	20
SAC-MP4SDQ/pDZ	1	−0.88	5.04	7.14	1.4
SAC-MP4SDQ/pTZ	1	−0.52	3.72	5.42	20
SAC-MP4/pDZ	1	−1.11	5.65	6.97	2.0
SAC-MP4/pTZ	1	−0.58	2.71	3.68	42
SAC-QCISD/pDZ	1	−0.58	4.35	6.48	2.3
SAC-CCD/pDZ	1	−0.88	5.49	8.23	2.5
SAC-CCSD/pDZ	1	−0.60	4.55	6.85	3.1
SAC-CCSD/pTZ	1	−0.31	3.97	5.37	31
SAC-CCSD(T)/pDZ	1	−0.58	4.46	6.00	3.9
SAC-CCSD(T)/pTZ	1	−0.18	1.99	2.92	105
EIB-MP2	2	−0.43	4.97	6.40	15
EIB-MP4D	2	−1.03	3.40	5.34	21
EIB-MP4SDQ	2	−0.77	3.58	4.87	21
EIB-MP4	2	−0.27	2.01	2.98	44
EIB-CCSD	2	−0.76	3.82	4.94	34
EIB-CCSD(T)	2	−0.19	1.49	1.97	109
MCSAC-MP4D/pDZ	2	−1.03	5.48	7.39	1.4
MCSAC-MP4D/pTZ	2	−0.95	3.62	5.26	20
MCSAC-MP4SDQ/pDZ	4	−0.93	5.12	7.05	1.4
MCSAC-MP4SDQ/pTZ	4	−0.30	2.26	3.16	20
MCSAC-MP4/pDZ	6	−0.53	4.49	5.96	2.0
MCSAC-MP4/pTZ	6	−0.28	2.23	3.15	42
MCSAC-CCSD/pDZ	2	−0.75	4.76	6.68	3.1
MCSAC-CCSD/pTZ	2	−0.56	2.87	3.70	31
MCSAC-CCSD(T)/pDZ	4	−0.15	4.07	5.56	3.9
MCSAC-CCSD(T)/pTZ	4	−0.04	1.67	2.31	105
MCCM-MP2	4	−0.71	3.47	4.38	15
MCCM-MP4D	6	−0.45	2.32	3.67	21
MCCM-MP4SDQ	8	−0.38	1.95	2.66	21
MCCM-MP4	10	0.05	1.33	1.73	44
MCCM-CCSD	6	−0.24	1.53	2.07	34
MCCM-CCSD(T)	8	0.02	0.81	1.12	109
MCCM-MP4D;MP2	5	−0.46	2.40	3.69	15
MCCM-MP4SDQ;MP2	6	−0.41	2.39	3.54	15
MCCM-MP4;MP2	7	−0.23	1.89	2.95	15
MCCM-CCSD;MP2	5	−0.25	1.94	2.74	17
MCCM-CCSD(T);MP2	6	0.05	1.01	1.40	17
MCCM-CCSD(T);MP2;HF+	7	0.10	0.95	1.29	18

[a] Computational cost is the mean CPU time on an IBM SP computer for the six largest molecules, H_2CCH_2, H_3CCH_3, H_3COH, H_2NNH_2, Si_2H_6, and CH_3SH, relative to the mean CPU time (66 s) for an MP2/pDZ calculation on the same computer for the same six molecules.

TABLE 4: Mean Signed Errors, Mean Unsigned Errors, and Rms Errors for IB Methods of Ref 10

method	parameters	MSE (kcal/mol)	MUE (kcal/mol)	RMSE (kcal/mol)	computational cost[a]
IB-MP2	2	3.75	6.37	7.97	15
IB-CCSD	2	−6.85	7.00	9.44	34
IB-CCSD(T)	2	−0.70	2.21	3.09	109

[a] Computational cost is the mean CPU time on an IBM SP computer for the six largest molecules, H_2CCH_2, H_3CCH_3, H_3COH, H_2NNH_2, Si_2H_6, and CH_3SH, relative to the mean CPU time (66 s) for an MP2/pDZ calculation on the same computer for the same six molecules.

Figure 4 gives an enlarged view of the nine recommended methods with a linear (rather than logarithmic) ordinate.

It is interesting to compare the parametrized and unparametrized methods. For this purpose we center attention on the unparametrized MP2/pTZ method (although the MP4/pTZ method has a slightly smaller mean unsigned error, it is 3 times more costly). The MCSAC-MP4SDQ/pDZ calculation gives a mean unsigned error of 5.12 kcal/mol with an average cost of 94 s, whereas MP2/pTZ gives a mean unsigned error of 6.74 kcal/mol with a cost of 893 s—the error for the unparametrized method is 1.3 times larger despite being 9.5 times more expensive. Similarly the MCCM-MP4;MP2 method gives a mean unsigned error 4.9 times smaller than MP2/pTZ at a cost that is only 1.1 times larger. Additional comparisons can be made for the remaining seven recommended methods, and these comparisons confirm the good performance of the MCCM methods.

As an additional check of accuracy and timing the MCCM methods we have calculated the zero-point-exclusive atomization energy of benzene with seven of the nine recommended methods

TABLE 5: Unsigned Errors (kcal/mol) of MCCM Methods Compared to Experiment for Benzene

method	unsigned error[a]	unsigned error per bond[b]	mean unsigned error per bond[c]	computational cost[d]
G2	6.97	0.58	0.53	1.00
MCSAC-MP4SDQ/pDZ	10.31	0.86	2.22	0.02
MCSAC-MP4/pDZ	11.82	0.99	1.95	0.05
MCCM-MP4;MP2	11.99	1.00	1.04	0.06
MCCM-CCSD;MP2	16.14	1.35	0.84	0.04
MCCM-CCSD(T);MP2	9.03	0.75	0.44	0.09
MCCM-CCSD(T);MP2;HF+	9.70	0.81	0.41	0.09
MCCM-CCSD	5.61	0.46	0.66	0.54

[a] Absolute value of error for benzene as compared to experiment. [b] Unsigned error for benzene divided by the total number of bonds, 12 bonds, in benzene. (Multiple bonds count as one bond.) [c] Sum of unsigned errors for the 49 values of D_e in the training set divided by the total number of bonds, 113 bonds, in the training set. (Double and triple bonds count as one bond.) [d] Computational cost is the CPU time for benzene on an IBM SP computer on a scale where the G2 calculation (3.0×10^5 seconds) is taken as 1 unit.

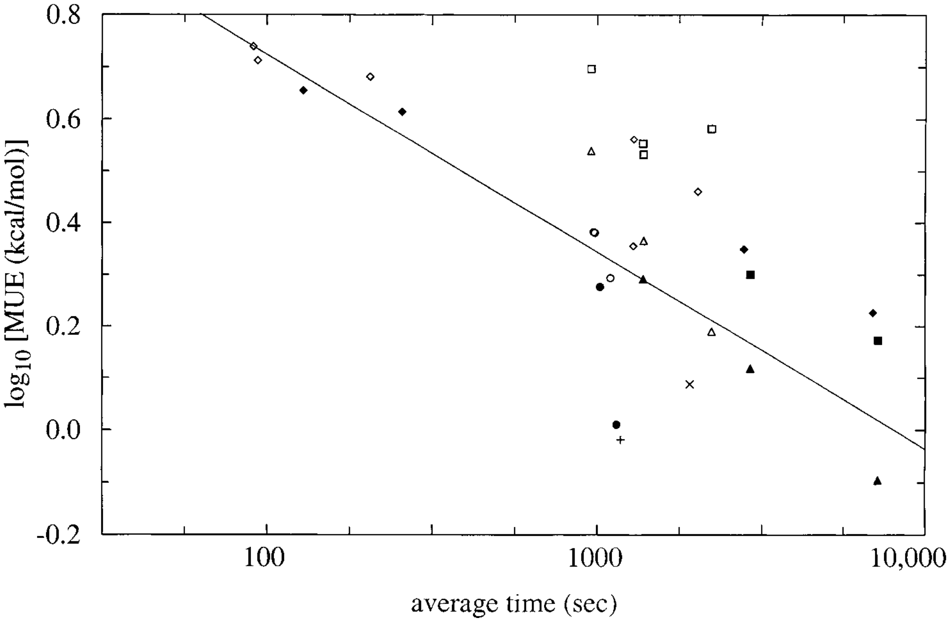

Figure 3. Logarithm of mean unsigned error (mean absolute error) as a function of cost for G2 and G2′ and all EIB, MCSAC, and MCCM methods. The cost is represented by the mean CPU time on an IBM SP computer for the six largest molecules, H_2CCH_2, H_3CCH_3, H_3COH, H_2NNH_2, Si_2H_6, and CH_3SH. Squares: EIB; diamonds: MCSAC; triangles: MCCM-L; circles: MCCM-L2;L1; +: MCCM-CCSD(T);MP2;HF+; ×: G2 and G2′. Methods including triple excitations are indicated by filled symbols, and those without triple excitations are indicated by hollow symbols.

and G2. The unsigned errors and unsigned errors per bond are given in Table 5. The experimentally based estimate of the zero-point-exclusive atomization energy of benzene, 1368.40 kcal/mol, was calculated from the experimental[30] value of $\Delta H_f{}^0{}_{298}$ by the method described in ref 9. As can be seen from Table 5, the error per bond for benzene for each of the methods is comparable to the average error per bond over the 49 molecule training set. (Benzene is not included in the training set.) Thus the mean error per bond does not systematically deteriorate with increasing system size. The computational costs for each method relative to G2 are given in Table 5. The G2 calculation has three correlation calculations (two MP4 calculations and one QCISD(T) calculation) with large basis sets that scale as N^7 where N is the number of basis functions and that dominate the computational cost for benzene. In contrast, the costs of the MCCM methods in Table 5 for benzene are dominated by calculations that either scale better than those in G2 or that scale the same as G2 but use a smaller basis set. The computational cost for an MCSAC-MP4SDQ/pDZ calculation is equivalent to that for an MP4SDQ/pDZ calculation that scales as N;[6] the cost of an MCSAC-MP4/pDZ calculation is equivalent to that for an MP4/pDZ calculation that scales as N^7 but that uses a smaller basis set than the MP4 calculations used in G2; and the MCCM-CCSD calculation is dominated by the CCSD/pTZ component which scales as N^6. The computational costs for the rest of the methods given in Table 5 are dominated by an MP2/pTZ calculation that scales as N^5. Therefore, because of the improved scaling or because of the use of a smaller basis set for the component of the calculation that scales as the highest power of N, we are able to significantly cut the cost of calculating larger molecules while maintaining very good accuracy.

7. Concluding Remarks

Motivated by the physical ideas of scaling correlation energy and extrapolating to an infinite basis set, we proposed an empirical multi-coefficient correlation method (MCCM) for *simultaneously* extrapolating both the various components of the correlation energy and the basis set. We have parametrized this idea for various levels of treating electron correlation with polarized double and triple-ζ basis sets and an augmented polarized double-ζ basis set. We obtain excellent agreement with experiment at reasonable cost without using the improperly dissociating high-level correction of Gaussian-1 and Gaussian-2 theory. Thus the method is well suited for calculating potential energy surfaces.

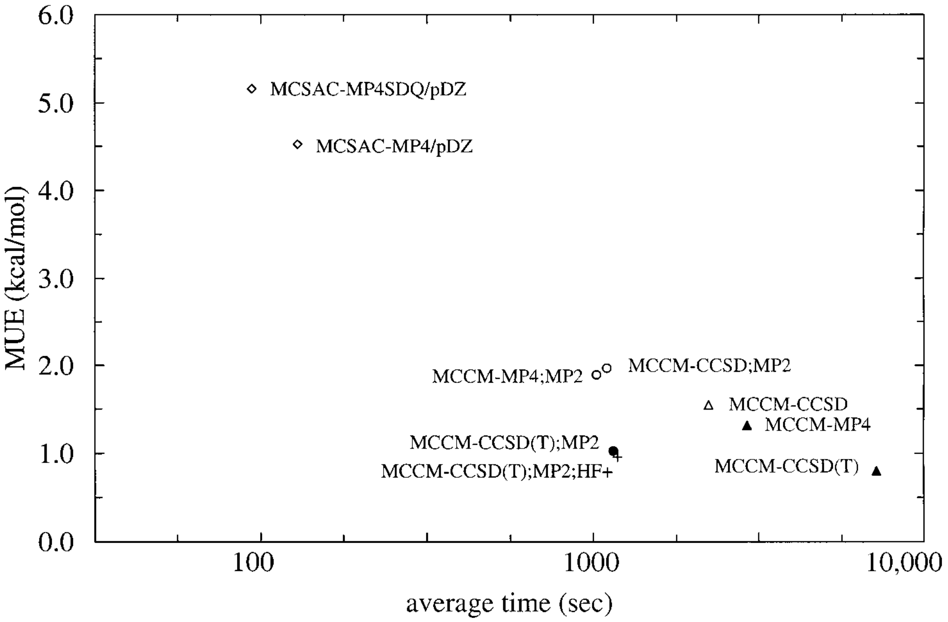

Figure 4. Mean unsigned error (mean absolute error) as a function of cost for the nine most highly recommended methods (solid circles). The cost measure is the same as for Figure 3.

Acknowledgment. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences. M.L.S. acknowledges a fellowship from the Ministerio de Educación y Cultura of Spain, which has supported her work at the University of Minnesota.

References and Notes

(1) Nyden, M. R.; Petersson, G. A. *J. Chem. Phys*. **1981**, *75*, 1843

(2) Brown, F. B.; Truhlar, D. G. *Chem. Phys. Lett*. **1985**, *117*, 307.

(3) Petersson, G. A. *ACS Symp. Ser*. **1998**, *677*, 237.

(4) Martin, J. M. L. *ACS Symp. Ser*. **1998**, *677*, 212.

(5) Blomberg, M. R. A.; Siegbahn, P. E. M. *ACS Symp. Ser*. **1998**, *677*, 197.

(6) Corchado, J. C.; Truhlar, D. G. *ACS Symp. Ser*. *712*, in press.

(7) Gordon, M.; Truhlar, D. G. *J. Am. Chem. Soc*. **1986**, *108*, 5412.

(8) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett*. **1995**, *234*, 64.

(9) Fast, P. L.; Corchado, J. C.; Sánchez, M. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 3139.

(10) Truhlar, D. G. *Chem. Phys. Lett*. **1998**, *294*, 45.

(11) Chuang, Y.-Y.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 651. Fast, P. L.; Sánchez, M. L.; Truhlar, D. G. *J. Chem. Phys.*, in press.

(12) Brown, F. B.; Truhlar, D. G. *Chem. Phys. Lett*. **1985**, *117*, 307.

(13) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys*. **1993**, *99*, 1914.

(14) Martin, J. M. L.; Taylor, P. R. *Chem. Phys. Lett*. **1996**, *248*, 336.

(15) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olson, J.; Wilson, A. K. *Chem. Phys. Lett*. **1998**, *286*, 243.

(16) Feller, D. *J. Chem. Phys*. **1992**, *96*, 6104.

(17) Feller, D.; Peterson, K. A. *J. Chem. Phys*. **1998**, *108*, 154.

(18) Dunning, T. H. Jr. *J. Chem. Phys*. **1989**, *90*, 1007.

(19) Woon, D. E.; Dunning, T. H. Jr. *J. Chem. Phys*. **1993**, *98*, 1358.

(20) Del Bene, J. E. *J. Phys. Chem*. **1993**, *97*, 107.

(21) Hehre, W. J.; Radom, L.; Schleyer, P.v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.

(22) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys*. **1982**, *76*, 1910.

(23) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett*. **1989**, *157*, 479.

(24) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys*. **1987**, *87*, 5968.

(25) *Gaussian94* (Revision E.2); Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Lahm, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. Gaussian, Inc.: Pittsburgh, PA, 1995.

(26) Woon, D. E.; Dunning, T. H. Jr. *J. Chem. Phys*. **1995**, *103*, 4572.

(27) Fast, P. L.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 3802.

(28) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. *J. Chem. Phys*. **1989**, *90*, 5622.

(29) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. *J. Chem. Phys*. **1991**, *94*, 7221.

(30) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Montgomery, J. A.; Frisch, M. J. *J. Chem. Phys*. **1998**, *109*, 10570.